AN INSTITIÚID
UM THAIGHDE
EACNAMAÍOCHTA
AGUS SÓISIALTA
ESRI ECONOMIC & SOCIAL
RESEARCH INSTITUTE

ESRI Working Paper No. 773

18 January 2024

Small Area Poverty Estimation by Conditional Monte Carlo

Niall Farrell[a,b*]

a) Economic and Social Research Institute, Dublin, Ireland
b) Trinity College Dublin, Dublin, Ireland

*Corresponding Author:
Dr Niall Farrell
Economic and Social Research Institute,
Whitaker Square, Sir John Rogerson's Quay,
Dublin, Ireland
Email: farreln7@tcd.ie

NF acknowledges funding from the Irish Environmental Protection Agency, the ESRI Energy Policy Research Centre and the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement number 743582. All remaining errors are the sole responsibility of the authors. Declarations of interest: none.

Abstract

Small area poverty estimates are important for social and economic policy, however the required data are often unavailable. This paper presents a Small Area Estimation (SAE) technique called Conditional Monte Carlo (CMC). CMC provides robust estimates of small area poverty rates, subject to fewer restrictive assumptions than existing methods. We present a theoretical derivation followed by a numerical validation. Using Mexican data, CMC replicates small area poverty rates with precision, successfully controlling for unobserved heterogeneity in the relationship between predictor and outcome variables through discriminate microdata sampling. CMC produces spatially-referenced microdata, providing a platform for agent-based modelling and microsimulation analysis.

1 Introduction

Small area poverty estimates are required to effectively target economic and social policy. They are the basis for federal funds apportionment in the United States (Ghosh et al., 1994; Tarozzi and Deaton, 2009), social policy decisions in many European countries (Simler, 2016) and poverty interventions in the developing world (Devarajan, 2013; Pokhriyal and Jacques, 2017). There is a growing requirement for small area estimates of economic and social outcomes. For instance, spatially-explicit poverty data will be required to target climate policy towards the microregions and socioeconomic groups that are most negatively affected by the impacts of climate change (Hallegatte and Rozenberg, 2017; Rao et al., 2017).

Poverty data at the small area level are often not readily available. This may be due to resource limitations or privacy concerns; survey microdata are frequently designed to be representative at the national or aggregated regional level. Small Area Estimation (SAE) methodologies exist to overcome this limitation, however common methodologies are subject to restrictive distributional assumptions (e.g. Elbers et al., 2003; Tarozzi and Deaton, 2009; Molina and Rao, 2010). This paper introduces a Conditional Monte Carlo (CMC) simulation-based SAE methodology. This method differs from many of those in the SAE literature as it provides poverty estimates by simulating a spatially representative population of households, building on methods from the spatial microsimulation literature (for a review, see O'Donoghue et al., 2014). This provides a number of advantages.

First, this method provides greater flexibility in the estimation procedure, relative to pre-existing methods. Many SAE procedures first estimate the conditional distribution of an outcome variable (such as income) using nationally or regionally-representative survey microdata. The small area distribution of this outcome variable is then imputed according to the census distribution of predictors. Under conditions of homogeneity in the relationship between outcome and predictor variables, these models give insight into the conditional expectation of a poverty outcome, accompanied by an estimate of precision. Discussed in Section 1, unobserved heterogeneity in this relationship at small area, sampling cluster or other spatial scales can bias estimates. Corrective procedures have been developed (e.g. Salvati et al., 2012), however many of these require user-defined parameters that are difficult to specify correctly. Unlike many SAE methods, the CMC procedure estimates small area poverty rates through selective sampling of households. Discriminate sampling of households for whom the relationship between outcome and predictor variables is shared with the small area of interest provides a less-restrictive means to overcome this restriction.

Second, this platform allows one to estimate the incidence of multiple indicators of deprivation and their likely coincidence within a household. Third, this method provides a platform with which one may apply microsimulation or agent-based methods to capture the complex interaction of policies, behaviours and economic shocks on the distribution of welfare, a policy priority identified in the literature (see Bourguignon and Spadaro, 2006; Cardaci, 2018; Farmer and Foley, 2009; Happe et al., 2008).

CMC is a population synthesis method building on those present in the spatial microsimulation literature. Many population synthesis methods to date employ computationally-intensive combinatorial optimisation methodologes (O'Donoghue et al., 2014). This yields a single snapshot simulation subject to sampling error. We employ a computationally efficient sampling method known as quota sampling (Farrell et al., 2013). This allows for multiple population snapshots through which one may estimate an expected rate of poverty, alongside a margin of error, providing estimates robust to sampling error.

This paper is structured as follows. Section 2 reviews the literature and motivates the development of the CMC methodology. This is followed by an overview of the theoretical problem in Section 3, where the CMC procedure is outlined. We present two validation exercises. First, we apply the CMC methodology to synthetic data in Section 4. A Monte Carlo experiment demonstrates the validity of the procedure in a general setting. In Section 5, we apply the method to an empirical case study. Using 2015 census data, we estimate small area poverty rates in Mexico and compare model performance against a known poverty distribution. We find that the method estimates small area poverty rates with precision. Section 6 offers some concluding comments.

2 Previous research and motivation

This section reviews the literature on poverty and small area estimation techniques, motivating the need for a robust microdata-based estimation procedure. Methods to estimate poverty incidence at the small area level may be categorised according to two research strands; Small Area Estimation (SAE) and Survey reweighting (SRW). SAE combines the micro-level power of survey data with the spatial information of census data in a multi-level modelling framework. SAE has been used in a wide range of applications. First proposed by researchers at the World Bank (Elbers et al., 2003), these methods have been employed in estimating the spatial distribution of poverty in over 60 countries (Elbers and van der Weide, 2014). Bedi et al. (2007) give an overview of applications in the developing world while Simler (2016) apply these methods to range of European countries. Further applications of note include the charting of spatial poverty incidence in Ecuador (Hentschel et al., 2000) and Vietnam (Minot, 2000).

Generally speaking, SAE procedures first estimate the conditional distribution of an outcome variable (such as income) using survey microdata. The spatial distribution of this outcome variable is then predicted according to the census distribution of predictors. Under specific distributional conditions, outlined below, these models give insight into the conditional expectation of a poverty outcome accompanied by an estimate of precision. This method was first proposed by Elbers et al. (2003). Often referred to as the ELL procedure (Tarozzi and Deaton, 2009), this is a simulation-based imputation procedure. The ELL model estimates a nested error-regression model. Random effects are controlled for at the sampling cluster level. While this controls for heterogeneity in the relationship between outcome and predictor variables at the sampling cluster level, it carries an implicit assumption that between-area variation in this relationship is negligible. Violation of this assumption will bias estimates. Tarozzi and Deaton (2009) argue that between-area variation in the relationship between predictor and outcome variables is likely in many circumstances, demonstrating the magnitude of potential bias.

The ELL model has been developed by subsequent methods in an attempt to overcome the limitations quantified by Tarozzi and Deaton (2009). Molina and Rao (2010) have created the Empirical Best Prediction (EBP) method, whilst Salvati et al. (2010) have introduced a method known as the M-Quintile (MQ) approach. The EBP method is similar to the ELL method but instead assumes heterogeneity in between-area variation, captured through a normally-distributed random effect at the area level. While the ELL model accounts for unobserved heterogeneity at the cluster level through a random effect, the EBP model requires the assumption of cluster homogeneity. Initially this method was sensitive to deviation from the assumption of Guassian random errors. Diallo (2014) and a paper by Elbers and van der Weide (2014) have developed augmented sampling procedures, such as a logarithmic or power transformation, to account for this. The M-quintile model, developed by

Chambers and Tzavidis (2006), is a further development that overcomes a number of the outlined deficiencies; it does not require any specific distributional assumptions, however correct implementation requires that the user specifies the parameters that should vary spatially, with no established diagnostic test in place to identify a definitive choice (Salvati et al., 2012).

To date, SAE methods have commonly been employed in the estimate of poverty as quantified by income, with some studies applying the method in the estimation of multidimensional poverty. Pham et al. (2020), use SAE methods to estimate subnational distributions of multidimensional poverty for Vietnam. These estimates are at regional level - small area distributions are not estimated. Pratesi et al. (2021) estimate the small area incidence of multidimensional educational poverty in Italy, with general poverty incidence not assessed.

Recently, SAE procedures have been augmented by the use of 'big data' sources, with Hall et al. (2023) providing a review. Of particular note is the work of Pokhriyal and Jacques (2017), who use data sources such as phone records and satellite imagery to predict important dimensions of poverty at the regional level. The spatial incidence of health, education, and standard of living have been estimated according to a Pearson correlation of 0.84–0.86. Machine Learning and Artificial Intelligence methods are also useful in this regard. Yeh et al. (2020), for instance, train machine learning algorithms to predict asset wealth across 20,000 African villages using publicly-available multispectral satellite imagery. Models explain 70% of the variation in village wealth for countries where the model was not trained.

SAE estimates such as these are useful but are restrictive with respect to the insight offered. First, these methods estimate headline poverty rates, one does not obtain insight into the micro-level distribution of poverty and its determinants. The small area covariance of poverty with unemployment, health outcomes or educational attainment, for instance, provides important information into the depth of deprivation in a given small area. This cannot be estimated with traditional SAE methods and is of particular importance given the move towards the characterisation of poverty on a multidimensional scale. There has been an increasing recognition that poverty and deprivation exists on many dimensions not wholly captured by income alone (e.g. Atkinson, 2003; Alkire and Foster, 2011; OPHI, 2013; Permanyer, 2014) and that poverty metrics should quantify incidence on multiple dimensions. Indeed, Sen (1999) has stated that 'the role of income and wealth. . . has to be integrated into a broader and fuller picture of success and deprivation' (Atkinson, 2003). A strong applied literature advocates for the use of multidimensional poverty indicators (Aaberge et al., 2019; Alkire and Foster, 2011; Atkinson, 2003; Curtis, 2018; Narayan-Parker and Patel, 2000; Permanyer, 2014; Sen, 1983, 1992, 2009; Santos and Villatoro, 2018; Stiglitz et al., 2009; Reddy and Pogge, 2002).

An estimation procedure that provides the distribution of relevant outcome variables, as opposed to the expected mean of a single metric, can provide insight into both the breadth and depth of poverty incidence, alongside a platform to investigate the dimensions most responsible for driving prevalence. This is valuable, given the importance of measuring poverty across multiple dimensions. In addition, understanding the distribution of relevant outcome variables - and their within-household coincidence - allows one to apply microsimulation or agent-based methods to capture the complex interaction of policies, behaviours and economic shocks. In doing so, this provides a platform to assess the effect that certain policies may have on the spatial distribution of welfare. This is a policy priority identified in the literature (see Bourguignon and Spadaro, 2006; Cardaci, 2018; Farmer and Foley, 2009; Happe et al., 2008).

Survey reweighting (SRW) methods overcome many of the aforementioned deficiencies, providing insight into micro-level distribution of determining variables. SRW reweighs survey microdata according to small area census totals. This provides insight into the coincidence of various outcomes while also providing a platform for spatially-explicit microsimulation and agent-based analyses (Tanton and Edwards, 2012; O'Donoghue et al., 2014; Rahman and Harding, 2016; Harding, 2017). This procedure is similar to SAE whereby a poverty estimate is produced using a microdata sample, reweighted conditional on known distributions at the small area level. While SAE methods estimate the conditional expectation of poverty rates using regression-based methods, SRW methods estimate a microdata distribution for each small area.

Many SRW methodologies exist and these may be categorised as either probabilistic and deterministic reweighting methods (for a full review of methods, see Tanton and Edwards, 2012; O'Donoghue et al., 2014; Rahman and Harding, 2016; Harding, 2017). These methods share two common deficiencies. First, all SRW techniques to date carry an implicit assumption of between-area homogeneity. Probabilistic reweighting methodologies such as simulated annealing operate by sampling indiscriminately from the survey data such that the simulated population of predictor variables corresponds to known distributions contained within census data at the small area level (for applications, see Ballas et al., 2006; Morrissey et al., 2014). As Tarozzi and Deaton (2009) discuss, the relationship between outcome and predictor variables is not necessarily homogeneous between areas and this can lead to a bias in estimation.

Second, common SRW methods present a single snapshot population from the conditional distribution and do not provide a robust measure of estimate precision. Probabilistic reweighting methodologies such as Simulated Annealing are computationally intensive (Rahman et al., 2010), and one cannot estimate multiple snapshots such that an estimate of precision may be approximated. Farrell et al. (2013) introduce the Quota Sampling method which is less computationally-intensive. However, this method provides a single snapshot and while applications to date involve the precision of this estimator has not been quantified.

'Deterministic' procedures, such as Iterative Proportional Fitting (IPF) create wholly synthetic tables of socioeconomic totals for a small area population, conditional on census distributions of predictor variables (for a full discussion, see Norman, 1999). Much research has applied this method, including Ballas and Clarke (2001); Ballas et al. (2005); Smith et al. (2009); Lovelace and Ballas (2013). The algorithm presents a number of deficiencies. The estimates produced are entirely synthetic and comprise a table of cross-tabulated population totals; insight into the within-household coincidence of outcome variables cannot be estimated, nor does the procedure provide a platform for agent-based or microsimulation analysis. In addition, the procedure is deterministic; the algorithm produces the same output each time the method is implemented (Lovelace and Ballas, 2013) and while estimates of precision have been traditionally therefore an estimate of precision is difficult to obtain. However, attempts have begun to incorporate an estimate of precision into these deterministic methods, with Whitworth et al. (2017) offering a method that borrows explanatory power from a multi-level regression. Rahman and Harding (2016) present a Markov-Chain Monte-Carlo method which provides this estimate of variance. The performance of these methods has not been validated to the author's knowledge.

The following section presents a survey reweighting methodology named Conditional Monte Carlo (CMC) which overcomes many of the outlined deficiencies of existing SRW methodologies. Discriminate sampling allows for area homoegenity to be overcome, while a bootstrapping methodology provides an estimate of precision. In doing so, the microdata format of the CMC output provides an approprate platform to (i) estimate spatial deprivation on a multidimensional scale, including estimates of the coincidence of various outcomes, and (ii) provide a foundation through which one can estimate the covariance of important socioeconomic outcomes or use as a platform for microsimulation or agent-based modelling.

3 The Theoretical Problem

3.1 Overview

Within Region C there is a small area i. The objective is to estimate the expected value of a statistical moment, W, which details the small area distribution of an outcome variable y for small area i, where $i \subset C$. y_i, where $y_i \in M_{H\times 1}(\mathbb{R})$, is a vector detailing the distribution of the outcome variable of interest for small area i. An example of such an outcome variable is the vector of household-level incomes or health status indicator for residents of small area i. W may be any statistical moment describing the distribution of the population in area i, such as the poverty index, crime index or the socioeconomic gradient of a health outcome. l_i is an estimate of the expected value of $W(y_i)$:

$$l_i = E[W(y_i)] \tag{1}$$

x_i is a matrix of K predictor variables correlated with the y_i outcome variable, where $x_i \in M_{H\times K}(\mathbb{R})$. For instance, should y_i be the distribution of income for small are i, x_i may be a matrix detailing the distribution of predictors such as education status or employment status. Using conditional expectations, we can transform the l_i estimator to the expected value for W, conditional on the observed X matrix equalling that observed for small area i, x_i:

$$l_i = E[W(y_i)|X = x_i]. \tag{2}$$

l_i may be estimated using Monte Carlo sampling:

$$l_i = E\Bigg[W(y^i)|X = x_i\Bigg] = \frac{1}{J}\sum_{j}^{J}\Bigg[W_j(y_i)|X_j = x^i\Bigg] \tag{3}$$

where X_j and $W_j(y^i)$ are the j^{th} Monte Carlo replication of the X covariate matrix and W estimate, respectively. Each Monte Carlo replication is a sample of households drawn from region C. The sample of households drawn from region C must be chosen such that the the X distribution of covariates corresponds to the known x_i distribution. Sampling a set of observations that exactly matches the x_i vector is a combinatorial optimisation problem. As Farrell et al. (2013), O'Donoghue et al. (2013) and Morrissey et al. (2014) discuss, combinatorial optimisation algorithms are often computationally intensive. This can preclude the Monte Carlo estimation outlined in equation (3). To overcome this constraint, the Conditional Monte Carlo estimator employs a computationally efficient rejection sampling procedure known as Quota Sampling developed by Farrell et al. (2013), which will now be outlined.

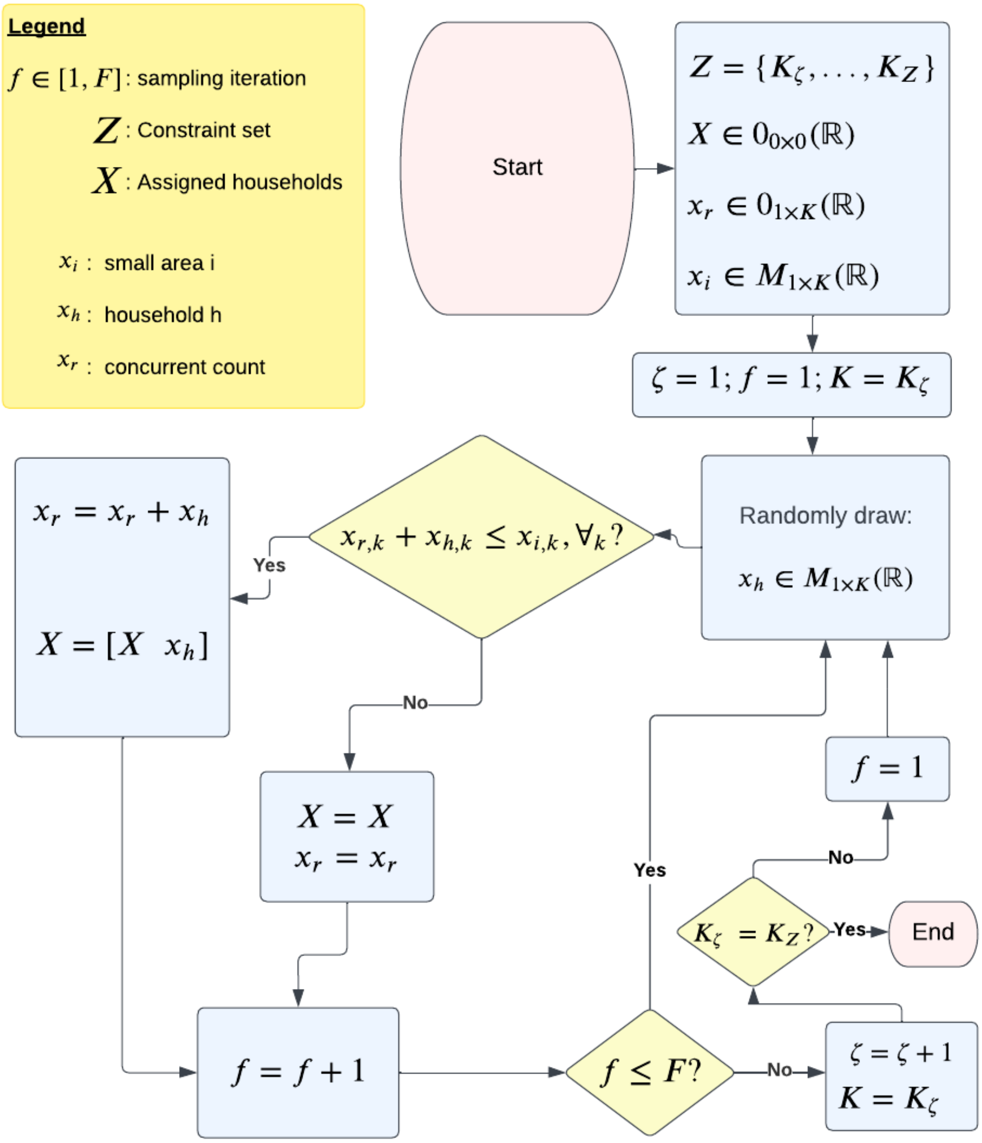
3.2 Rejection sampling

The rejection sampling procedure draws on the quota sampling algorithm first introduced by Farrell et al. (2013) and O'Donoghue et al. (2013). This procedure will now be outlined, with a full discussion in Farrell et al. (2013). As outlined in the previous section, the objective is to sample an X matrix which details the household-level covariate distribution from the sampled population, where $X \in M_{H\times K}(\mathbb{R})$. Each h row of the X matrix represents the household-level distribution of key predictor variables, detailed according to the K columns. The X matrix is therefore an appending of x_h row vectors.

The Quota sampling procedure is the process by which these row vectors are sampled and appended such that $X = x_i$. This involves sampling H observations (e.g. households) from a microdata population C such that X vector outlining the covariate distribution for sampled households approximates the known x_i distribution for small area i:

$$X \approx x_i, \tag{4}$$

Figure 1: Quota sampling rejection sampling procedure



Quota Sampling procedure first introduced by Farrell et al. (2013). Please see Farrell et al. (2013) for a full discussion of this procedure.

The practical implementation of this quota sampling concept is illustrated in Figure 1 and discussed at length in Farrell et al. (2013). As previously discussed, households are sampled according to K covariates. Z is the set of K covariate specifications which may be employed in the estimation process, where K_ζ represents the concurrent covariate set (i.e. $Z = K_\zeta, ..., K_Z)$[1]. The specification of multiple covariate sets is required to overcome practical limitations to convergence. As Farrell et al. (2013) outline, the sampling procedure does not replace households that have already been assigned to small area i. This improves efficiency relative to other combinatorial optimisation algorithms such as simulated annealing. While this brings computational efficiency, facilitating the implementation of the Conditional Monte Carlo procedure, there are practical limitations to convergence. Given the intra-household distribution of covariates, the final household to be assigned may be required to have characteristics that do not exist to perfectly satisfy the distribution represented by a given set of K covariates.[2] While this introduces some additional noise to the estimate, this should not bias the estimate under the assumption that the households assigned under all sets of constraints are randomly allocated, conditional on $X = x_i$. This hypothesis is tested in Section 4.

The procedure operates as follows. We first specify a number of vectors to facilitate algorithm operation. The algorithm allocates households to small area i one at a time. An x_r vector logs the concurrent count of attributes for the set of households that have been already assigned to small area i at a given point in algorithm operation. For a given K set of predictor variables, x_r is a vector of zeroes at the initiation of each algorithm loop: $x_r \in 0_{1\times K}(\mathbb{R})$. At initiation, X is an empty matrix.

Within each algorithm loop, the N set of households in the candidate sample dataset are sorted randomly. Household h is extracted from the N sample population, with an x_h vector of covariate attributes, where $x_h \in M_{1\times K}(\mathbb{R})$. The algorithm then

[1] For simplicity, we refer to the set of constraints in each simulation iteration as K, where $K = K_\zeta$. See Figure 1

[2] For instance, we may be controlling for occupation and education status, with the final household requiring a scientist without a third level qualification. This may not be present in the dataset.

compares the concurrent count in addition to the addition of household h to the constraint counts for small area i. Household h is allocated to small area i if an elementwise addition of vectors $x_h + x_r$ produces a value less than or equal to the corresponding elements from x_i, for all k vector elements:

$$x_{r,k} + x_{h,k} \leq x_{i,k}, \forall_k \tag{5}$$

If equation (5) is satisfied, household h is allocated to small area i. Vector x_r is updated to reflect the new quota counts. Matrix X is then updated, by concatenating the x_h values with the pre-existing X matrix, which contains previously assigned households.

If Equation (5) is not satisfied, then household h is not allocated to x_r. The 'quota' of socioeconomic attributes for at least one attribute has been exceeded. The algorithm continues without updating the X matrix or the x_r vector and household $h + 1$ is evaluated. Households are evaluated consecutively until $x_r \approx x_i$.

Should the algorithm fail to allocate households a set F number of times, then a subsequent K constraint set is invoked, whereby $K = K_{\zeta+1}$. $K_{\zeta+1}$ corresponds to fewer constraining variables. This imposes fewer constraints on the statistical match than K_ζ, facilitating the allocation of additional households to small area i.[3] The algorithm continues in this fashion until the required number of households have been allocated. Once the X matrix of household attributes is constructed, this is matched with the corresponding y_i outcome variable. The specification of the small area poverty distribution is then complete.

[3] Without such a step, the intra-household distribution of constraints may result in the algorithm failing to allocate additional observations. Ex-ante, one would expect that allocating additional households subject to fewer constraints should improve the precision of the estimate relative to an estimate based on fewer households. We test this hypothesis in Section 4.

3.3 Unbiased inference and area homogeneity

Having outlined the CMC estimation procedure, and reviewing the quota sampling subroutine first introduced by Farrell et al. (2013), the next step is to consider the assumptions that are important for unbiased inference. Small area i is a subset of region C. Elbers et al. (2003) and Tarozzi and Deaton (2009) show that estimating $E[W(y_i)|X = x_i]$ can yield unbiased results should the covariance between $W(y_i)$ and x remains consistent throughout region C. In other words, there must be homogeneity in the relationship between predictor and outcome variables for households located in all small areas within region C. If small area-specific effects are present, then this relationship will not hold. Formally, the requirement for area homogeneity may be characterised as:

$$E[W(y_i)|X = x_i], h \in C = E[W(y_i)|X = x_i], h \in i \tag{6}$$

Equation (6) implies that if there is regional variability, such as a lack of access to certain resources in certain locations, indiscriminate geographical sampling may fail to capture the full degree of spatial heterogeneity and estimates will be biased. The CMC sampling algorithm can overcome this constraint in many circumstances by discriminately sampling from regions where area homogeneity holds. This is explored in Section 5.

4 Monte Carlo Experiments

This section is the first step in evaluating the Conditional Monte Carlo procedure. First, we carry out a set of Monte Carlo experiments to test the performance of the model using a synthetic data generation process. The focus of this process is to test the impact that estimation choices have on model performance.

A number of propositions are tested. First, we validate the procedure as an unbiased estimator of a statistical moment of interest. Second, we test the added precision introduced by the sequential quota sampling procedure.[4] Third, we evaluate the impact of unobserved spatial heterogeneity on the precision of the estimated statistical moment. We follow Tarozzi and Deaton (2009) and introduce an area-specific fixed effect to capture this. Finally, we test sensitivity to the number of Monte Carlo iterations chosen, informing practical application. We consider the precision and bias of the estimate when measuring estimator performance.

4.1 Data Generation Process

The Monte Carlo experiment proceeds as follows. We adapt the Data Generation Process of Tarozzi and Deaton (2009), to simulate a set of households. Each household belongs to a small area i, where the small area may have a region-specific fixed effect. Two covariate variables, $x_{1,ih}$ and $x_{2,ih}$ predict the y_{ih} poverty indicator for household h. u_{ih} denotes the household-specific error which comprises an observation-specific error ϵ_{ih} and a small area-specific error η_i. We assume the constant β_0 is equal to 10. Given these parameters, y_{ih} is simulated according to the following Data Generation Process (DGP):

$$\begin{aligned} y_{ih} &= \beta_0 + \beta_1 x_{1,ih} + \beta_2 x_{2,ih} + u_{ih} \\ ... &= 10 + x_{1,ih} + x_{2,ih} + \psi \cdot \eta_i + \epsilon_{ih} \end{aligned}$$

[4] As outlined in Section 3, a $\zeta = 1, ..., Z$ set of constraint specifications may be specified. We compare estimator performance under a single set of constraints to performance when there are multiple sets of predictor variables.

where:

$$
\begin{aligned}
x_{1,ih} = 0.5 * (5 + z_{1,i}w_{1,ih} + z_{2,i}),\ w_{1,ih} \sim N(0,1),\\
x_{2,ih} = 0.5 * (5 + z_{2,i}w_{2,ih} + z_{4,i}),\ w_{2,ih} \sim N(0,1),\\
z_{1,i}, z_{2,i}, z_{3,i}, z_{4,i} \sim U(0,1),\\
z_{1,i} \perp z_{2,i} \perp z_{3,i} \perp z_{4,i},\\
\eta_i \sim N(0, .01), \epsilon_{ih} \sim N(0, \sigma^2(x))\\
\sigma^2(x) = \frac{e^{\alpha_1 x + \alpha_2 x^2}}{1 + e^{\alpha_1 x + \alpha_2 x^2}}.
\end{aligned}
$$

In addition, $\alpha_1 = 0.05$ and $\alpha_2 = 0.01$. Following Tarozzi and Deaton (2009), the idiosyncratic errors ϵ_{ih} are assumed to be heteroskedastic. ψ denotes scaling parameter for the small area-specific error and is assumed to be 2 in our baseline model specification. We vary this to test the sensitivity of estimation performance to unobserved spatial heterogeneity.

The validation procedure proceeds as follows. First, we simulate census distributions for each variable according to the outlined DGP. We simulate a population of 4,500 households divided into 15 small areas of 300 households each. Many census datasets present their data according to discrete categorisations for relevant variables. To implement the CMC procedure five discrete categorisations for both $x_{1,ih}$ and $x_{2,ih}$ variables are specified.[5] As with the $x_{1,ih}$ and $x_{2,ih}$ variables, we match according to a discrete $\psi \cdot \eta_i$ variable, which is split into 7 discrete categories.[6] The second step is

[5] These categories are as follows: category 1: $x < 4$; category 2: $4 < x \leq 5$; category 3: $5 < x \leq 6$; category 4: $6 < x \leq 7$; category 5: $7 < x$

[6] In the baseline specification, where $\psi = 2$, the $\psi \cdot \eta_i$ discrete categorisations take the following values: category 1: $\psi \cdot \eta_i \leq -0.644$; category 2: $-0.644 < \psi \cdot \eta_i \leq -0.275$; category 3: $-0.275 < \psi \cdot \eta_i \leq -0.145$; category 4: $-0.145 < \psi \cdot \eta_i \leq 0.025$; category 5: $0.025 < \psi \cdot \eta_i \leq 0.248$; category 6: $0.248 < \psi \cdot \eta_i \leq 0.420$; category 7: $0.420 < \psi \cdot \eta_i$

to generate a psuedo survey, from which the CMC estimator may select observations to emulate the identified census distributions. we generate a 10% survey sample of 450 households, from which we sample using the CMC algorithm.

The final step is to implement the CMC procedure. The CMC estimation procedure selects observations from the survey subsample to emulate the known discrete census distributions, calculated earlier. The baseline CMC specification has two Z constraint sets and 250 Monte Carlo iterations. Should the algorithm fail to allocate 10 households in sequence, the algorithm moves to the subsequent constraint set. Following Tarozzi and Deaton (2009), all Monte Carlo replications use the same artificial census population, which is therefore treated as non-random. We test the sensitivity to these parameterizations.

For each household, the y_{ih} variable may be interpreted as a metric of income or well-being. We follow Tarozzi and Deaton (2009) and estimate the head count poverty ratio as the statistical moment of interest, calculated as the percentage of the population living below a given poverty line, $P(y_{ch})$. We choose $y_{ih} = 14$ as the poverty line. In our baseline specification this yields a headcount poverty rate of around 41%, a benchmark that is suitable for evaluating model performance.

We calculate a number of statistics to give insight into estimation performance. First, we inspect any bias in the algorithm by calculating relative bias as:

$$\frac{\sum^{J}}{J} \cdot \frac{(\hat{y_{ihj}} - y_{ih})}{y_{ih}}, \tag{7}$$

where y_{ih} is the true value of the welfare measure, and $\hat{y_{ihj}}$ is the estimate obtained in the $J - th$ Monte Carlo replication. Second, we calculate an estimate of precision using Root Mean Square Error (RMSE). The RMSE is estimated as:

$$\sqrt{\frac{\sum^{J}(\hat{y_{ihj}} - y_{ih})}{J}} \tag{8}$$

Finally, we calculate the fraction of simulation replications which lie within one or two standard deviations of the population mean (i.e. $\mu + 1\sigma$ & $\mu + 2\sigma$). According to the empirical rule, 68% of estimated means should lie within one standard deviation of the mean, whilst 95% of estimated means should lie within two standard deviations of the mean.

4.2 Monte Carlo experiment results

The first proposition we wish to test is the performance of the CMC algorithm in accurately estimating the poverty headcount for a fictional small area population. We do so according to the data generation process previously outlined with baseline parameter values. Outlined in Table 1, we include a number of Z constraint set specifications to give insight into performance under various constraint specifications. Each specification varies the presence of constraints in step two ($\zeta = 2$) in the simulation procedure. We first consider a scenario where all constraints are present in both steps (Specification A), then we test the performance of the algorithm when we remove one constraint in step two (Specifications B-D).

Table 1: CMC constraint specifications

Specification	Constraints	
	$\zeta = 1$	$\zeta = 2$
A (All)	x_1, x_2, $\psi \cdot \eta_i$	x_1; x_2; $\psi \cdot \eta_i$
B (No x_1 when $\zeta = 2$)	x_1; x_2, $\psi \cdot \eta_i$	x_2; $\psi \cdot \eta_i$
C (No x_2 when $\zeta = 2$)	x_1; x_2; $\psi \cdot \eta_i$	x_1; $\psi \cdot \eta_i$
D (No $\psi \cdot \eta_i$ when $\zeta = 2$)	x_1; x_2; $\psi \cdot \eta_i$	x_1; x_2
E (No $\psi \cdot \eta_i$ when $\zeta = 1$ & $\zeta = 2$)	x_1; x_2;	x_2

Table defines constraint specifications employed in Table 2. Each specification varies the presence of constraints in step two ($\zeta = 2$) in the simulation procedure. We first consider a scenario where all constraints are present in both steps (Specification A), then we test the performance of the algorithm when we remove one constraint in step two (Specifications B-D).

Table 2 reports the results of this examination. For all constraint specifications, bias and RMSE is small relative to the true value being estimated. Specifications B and C perform better than specifications A and D. Relative to specifications B and C, specification A presents a greater bias and a marginally lower number of simulation iterations which fall within the interval predicted by the empirical rule. This is because the removal of constraints in specifications B and C allows for further households to be added and closer convergence with the true poverty rate. As results show, this procedure reduced bias, although there is an insignificant effect on RMSE.

Simulation performance is best when either x_1 or x_2 are removed from the second simulation step in Specifications B and C. The proportion of simulation iterations that fall within the intervals $\mu + 1\sigma$ & $\mu + 2\sigma$ converges on those predicted by the empirical rule. Specification D demonstrates a poorer performance relative to specifications B and C. The performance is comparable to Specification A. This is because removal of $\psi \cdot \eta_i$ results in a loss of control of spatial heterogeneity in the estimation procedure.

This has a greater impact than the removal of either x_1 or x_2.

In addition to the findings of Table 2, we wish to explore the extent with which unobserved spatial heterogeneity may bias CMC estimates. In the results of Table 2, we control for spatial heterogeneity through the explicit incorporation of the $\psi \cdot \eta_i$ parameter in the CMC estimation. To test for the influence of unobserved heterogeneity, we apply the CMC algorithm to Specification E in Table 1, whereby we do not control for the small area heterogeneity component. We test the varying influence of spatial heterogeneity by varying the $\psi \cdot \eta_i$ parameter.

These experimental findings are demonstrated in Table 3which shows that the introduction of uncontrolled spatial heterogeneity results in poor model performance. A number of ψ parameter specifications are chosen, ranging from 0.1 to 1. We see that small degrees of spatial variation, where $\psi \leq 0.2$, result in negligible impacts on model performance. Indeed, model performance is marginally better than the results of Table 2, due to less variation in parameter specification. However, once $\psi \geq 0.5$, the spatial heterogeneity becomes a large predictor of the outcome variable and model performance diminishes significantly. When $\psi = 1.0$, the model is unable to predict income. This result corresponds to that of Tarozzi and Deaton (2009) show, incorporation of regional heterogeneity is of importance for adequate performance of the CMC algorithm.

The final analysis of this section is to consider model sensitivity to further model specification choices. In particular, we wish to test performance relative to the number of simulation iterations. The results of this exercise are shown in Table 4, where we see that once the algorithm is applied to carry out 100 iterations, performance converges with that expected.

Table 2: Comparing CMC specification in estimating $P_0(14)$ poverty rate

Specification	Detail	Bias	RMSE	Interval	
				$\mu + 2\sigma$	$\mu + 1\sigma$
A	All constraints	-0.045	0.106	0.924	0.648
B	no x_1 when $\zeta = 2$	-0.023	0.107	0.948	0.696
C	no x_2 when $\zeta = 2$	-0.020	0.111	0.960	0.676
D	no $\psi \cdot \eta_i$ when $\zeta = 2$	-0.037	0.103	0.936	0.652

Results of CMC algorithm employed according to baseline specification with 250 Monte Carlo iterations. $P_0(14)$ represent the head count poverty rate for poverty line of 14. According to the empirical rule, 68% of estimated means should lie within one standard deviation of the mean (i.e. $\mu + 1\sigma$), whilst 95% of estimated means should lie within two standard deviations of the mean (i.e. $\mu + 2\sigma$).

Table 3: CMC algorithm performance: Influence of unobserved spatial heterogeneity when estimating $P_0(14)$ poverty rate

η_i	True value	Bias	RMSE	Interval	
				$\mu + 2\sigma$	$\mu + 1\sigma$
0.1	7.000	0.003	0.027	0.956	0.644
0.2	7.667	-0.004	0.027	0.968	0.736
0.5	11.000	-0.037	0.046	0.724	0.288
1	21.667	-0.140	0.143	0.008	0.004

Results of CMC algorithm employed according to baseline specification with 250 Monte Carlo iterations. $P_0(14)$ represents head count poverty rate for poverty lines of 14. According to the empirical rule, 68% of estimated means should lie within one standard deviation of the mean (i.e. $\mu + 1\sigma$), whilst 95% of estimated means should lie within two standard deviations of the mean (i.e. $\mu + 2\sigma$).

Table 4: CMC algorithm performance: Comparing Monte Carlo iteration specifications when estimating $P_0(14)$ headcount poverty rate

Monte Carlo iterations	Bias	RMSE	Interval	
			$\mu + 2\sigma$	$\mu + 1\sigma$
10	-0.049	0.107	0.900	0.700
50	-0.034	0.095	0.940	0.640
100	-0.013	0.108	0.980	0.590
250	-0.023	0.107	0.948	0.696
500	-0.019	0.104	0.954	0.670
1000	-0.018	0.102	0.953	0.673

Results of CMC algorithm employed according to baseline specification with 250 Monte Carlo iterations. $P_0(14)$ represents head count poverty rates for poverty line of 14. According to the empirical rule, 68% of estimated means should lie within one standard deviation of the mean (i.e. $\mu + 1\sigma$), whilst 95% of estimated means should lie within two standard deviations of the mean (i.e. $\mu + 2\sigma$).

5 Empirical application

The second stage of the CMC validation process is an empirical application. We construct a quasi-validation procedure using a Mexican case study. The focus of this step is to demonstrate an empirical application while exploring the performance of discriminate sampling to capture unobserved spatial heterogeneity.

5.1 Data and Methods

We use data from the 2015 census of Mexico Integrated Public Use Micro Sample (IPUMS). IPUMS is a 9.5% random extract from the 2015 Census of Mexico, totalling 11,344,365 observations. These data are chosen for two reasons. First, the Mexican census exhibits considerable between-area heterogeneity, as demonstrated by Tarozzi and Deaton (2009), and we wish to evaluate the performance of the CMC estimation procedure in capturing this variation. Second, IPUMS data allows for the estimation of spatially-referenced multidimensional poverty indicators. This demonstrates the analytical insight possible with the CMC microdata-based estimation procedure.

The validation procedure proceeds as follows. First, a psuedo-census is created. The 9.5% census sample is replicated according to household weights, with each household replicated accordingly. This provides a psuedo population for all of Mexico. Each state is subdivided into municipios (municipalities), which we take to be the small areas of our analysis. For each small area/municipio i, we calculate the distribution of the x_i vector of predictor variables associated with the y_i outcome variable and the W population moment of interest.

A psuedo survey is then constructed to emulate the Small Area Estimation procedure. The psuedo-survey is calculated as a random sample of 2000 households from the entire population census. Following the findings of Section 4, we implement the CMC procedure with 100 Monte Carlo replications. From these 100 simulation iterations, estimates of poverty are calculated for each small area.

Tarozzi and Deaton (2009) have shown that unobserved between-area heterogeneity exists between Mexican municipailities, driving bias in estimates using standard SAE techniques. As section 3 has outlined, we postulate that this bias may be overcome through discriminate sampling from regions where unobserved factors driving the population moment of interest are homogenous between candidate households. We test this hypothesis in the following way. Mexico is divided into states and, within each state, into municipios (municipalities). When implementing the CMC estimation procedure for a given municipio, candidate households comprise those sourced from the same state as that municipio. This carries an implicit assumption that there is a homogenous relationship between outcome and predictor variables for all municipios within a given state.

5.2 Metrics of analysis

We use metrics to measure estimator precision and to quantify poverty incidence by small area. In addition, we wish to investigate the socioeconomic drivers most responsible for observed poverty rates. We use the concentration index to illustrate this. The metrics employed for each of these objectives will now be outlined.

5.2.1 Metrics of estimator precision

Estimate precision is quantified using 3 metrics: correlation; average small area RMSE; and average small area relative bias, as outlined in Section 4. While Section 4 considered the proportion of simulation iterations that fall within thresholds of the population mean to identify estimate bias, we wish to complement this insight with an estimate of estimator precision. As such, we identify the proportion of small areas for whom the actual poverty value falls within s standard deviations of the simulated mean. In other words, we estimate the proportion of small areas where the actual poverty value is contained within reported confidence intervals. We consider two confidence interval thresholds: $s = 2$ (i.e. actual poverty rate falls within simulated range 95% of the time) and $s = 3$ (i.e. actual poverty rate falls within simulated range 99% of the time).

5.2.2 Measuring poverty using a Multidimensional Poverty Index

For this application, we calculate household welfare according to multidimensional poverty metric. We define multidimensional poverty using the MPI for Latin America index (MPI-LA) proposed by Santos and Villatoro (2018). Reviewed in Section 1, a strong applied literature advocates for the use of poverty indicators that characterise poverty on many dimensions (Aaberge et al., 2019; Alkire and Foster, 2011; Atkinson, 2003; Curtis, 2018; Narayan-Parker and Patel, 2000; Permanyer, 2014; Sen, 1983, 1992, 2009; Santos and Villatoro, 2018; Stiglitz et al., 2009; Reddy and Pogge, 2002), while the UN have formally acknowledged multidimensional poverty in the achievement of their sustainable development goals.[7] While the definition of poverty on multiple dimensions may seem unnecessarily complex for the neoclassical economist[8], markets do not necessarily function optimally in many circumstances and not all services are acquired through market transactions. This is particularly true in a developing world context. Participatory studies back up the importance of multiple dimensions, showing that the poor themselves describe their deprivations in terms beyond a lack of income (Naraya et al., 2000; UNDP, 2013; Santos and Villatoro, 2018).

There are a number of MPI metrics. The most common being the Alkire-Foster (AF) (Alkire and Foster, 2011) method which, among other traits, generates a poverty measure that is sensitive to the depth of deprivation (i.e. the number of dimensions through which deprivation is experienced). This allows for deprivation to be broken down according to dimension and the joint distribution of various deprivations to be estimated. Critical dimensions of housing, health, education, material deprivation and employment are accounted for (Burchi et al., 2019). This method is applied at a global scale with the Oxford Poverty and Human Development Institute (OPHDI)

[7] Target 1.2 of the Sustainable Development Goals extends poverty beyond the dimension of income and calls for a reduction in ‘poverty in all its dimensions according to national definitions’ (UN, 2015).

[8] The ability to attain a given standard of living is a function of resources at one’s disposal. Resources, in turn, are acquired through market transactions. Conditional on well-functioning markets, income or expenditure may indeed provide sufficient metrics of access to marketable resources.

and the UNDP publishing an annual report of the Global MPI (e.g. Alkire et al., 2019).

Variants of the AF procedure exist. Santos and Villatoro (2018) have developed a variant of the AF method for Latin American countries. Dubbed MPI-LA, this better captures poverty in this region for a number of reasons. Firstly, it borrows from basic needs, capability and rights-based approaches in the incorporation of indicators. Secondly, it combines monetary and non-monetary indicators and aligns deprivation cutoffs with Latin American living standards. Furthermore, it accounts for deprivations in employment, social protection and schooling. These are indicators unaccounted for in metrics of unmet basic need, heretofore commonly employed (Santos and Villatoro, 2018).[9]

The MPI-LA poverty metric is contructed in the following way. There are two constituent components: MPI poverty headcount (PH) and MPI poverty depth (PI). Let $x_h \in \mathbb{R}_+$ be the vector of predictor variables for household h. For each predictive variable k, the household is considered deprived in that variable should the observed $x_{h,k}$ value be less than a γ_k cutoff, where $k \in \zeta$: $\mathbb{1} \cdot (x_{h,k} < \gamma_k), k \in \zeta$

To calculate the MPI-LA poverty index, one must first calculate whether a household is in poverty. Denoting poverty status as y_h, the outcome variable of interest is calculated as a weighted sum of each $\mathbb{1} \cdot (x_{h,k} < \gamma_k), k \in \zeta$ poverty indicator:

$$y_h = \sum_k {}_K w_k \cdot x_{h,k}, k \in \zeta \tag{9}$$

where w_k is the weight for predictor variable k

The poverty headcount for small area i (PH_i) is defined as the proportion of households within small area i for whom the y_h indicator is greater than the *theta*

[9] For further information on the MPI-LA approach, see Santos and Villatoro (2018).

poverty cutoff:

$$PH_i = \frac{\sum_H \mathbb{1} \cdot (y_h > \theta)}{H}, \tag{10}$$

where H is the number of households in small area i. Poverty depth is defined as the average deprivation score conditional on the deprivation score being greater than the poverty threshold θ:

$$PI_i = \frac{\sum_H y_h \cdot \mathbb{1} \cdot (y_h > \theta)}{\sum_H \mathbb{1} \cdot (y_h > \theta)}, \tag{11}$$

the MPI-LA poverty indicator is then the product of poverty intensity and the poverty headcount:

$$(''MPI - LA'')_i = PH_i \cdot PI_i, \tag{12}$$

Following the definition of MPI-LA, the x^k dimensions characterised by Santos and Villatoro (2018) are adopted for this paper in the characterisation of multidimensional poverty and outlined in Table 5.

For the CMC simulation procedure, $Z = 3$ constraint sets are employed. 13 constraints are chosen that correlate with the MPI-LA dimensions of Table 5. All 13 constraints are present when $\zeta = 1$. 7 constraints are removed when $\zeta = 2$. These are removed as they are highly correlated with variables retained in the simulation, and therefore reduce simulation complexity while retaining strong predictive power. The $\zeta = 3$ simulation iteration corresponds to a close to random allocation of remaining households, with only precarious roof controlled for. In many simulation iterations, the required number of households have been allocated and this stage is skipped.

Table 5: Multidimensional poverty dimensions and indicators

Dimension (x_k)	Detail	Weight (w_k)
Housing		
Housing materials	Dirt floor, precarious roof or wall materials[1]	7.40%
People per room	3+ per room	7.40%
Housing tenure	Illegally occupied, ceded or borrowed house	7.40%
Basic services		
Improved water source	Not piped to house	7.40%
Improved sanitation	No unshared plumbed toilet in dwelling	7.40%
Energy	No electricity or cook with wood/coal/dung	7.40%
Living standards		
Monetary resources	Insufficient income for food & nonfood needs	14.8%
Durable goods	No car, refridgerator, washing machine.	7.40%
Education		
Children's school attendance	Child/adolescent not attending school	7.40%
Schooling gap	Child/adolescent 2+ yrs delayed schooling grade	7.40%
Adult school achievement	Nobody 18+ w/ primary & secondary educ	7.40%
Employment and social protection		
Employment	1+ member employed	7.40%
Social protection	Nobody w/ social protection[2]	3.70%
Poverty cutoff η		25

[1] Waste, carboard, tin, cane palm, straw, other

[2] health insurance, contributing to social security system, receiving pension/retirement income

Table 6: Constraints used in CMC simulation: empirical application

	Constraint used for each ζ		
Description	$\zeta = 1$	$\zeta = 2$	$\zeta = 3$
Housing			
Precarious roof in dwelling[1]	Yes	Yes	Yes
3+ people per room	Yes	Yes	No
Urban location	Yes	Yes	No
Dirt floor in dwelling	Yes	No	No
Illegally occupied, ceded or borrowed house	Yes	No	No
Basic services			
Water not piped to house	Yes	No	No
No unshared plumbed toilet in dwelling	Yes	Yes	No
Education			
Nobody 18+ w/ primary & secondary education	Yes	No	No
Child/adolescent not attending school	Yes	Yes	No
Child/adolescent 2+ yrs delayed schooling grade	Yes	Yes	No
Employment and social protection			
1+ household members employed	Yes	Yes	No
Any adult/child did not eat a meal due to insufficient money	Yes	Yes	No
Nobody w/ social protection[2]	Yes	No	No

[1] Waste, carboard, tin, cane palm, straw, other

[2] health insurance, contributing to social security system, receiving pension/retirement income

5.2.3 Investigating the drivers of poverty using the Concentration Index

We wish to apply the unit-level estimation procedure to investigate the drivers of poverty incidence. We demonstrate the utility of such insight through the the concentration index (CI), a metric which quantifies the socioeconomic gradient of an outcome variable of interest along the income distribution. This gives insight into both the small area distribution of a statistical moment, such as the multidimensional poverty rate, but also the extent with which this statistical moment is correlated with a given determining variable along the income distribution. The CI is easily quantified and is measured between the ranges -1 and 1, with zero representing perfect equality. The concentration index may be calculated as:

$$CI = \frac{2}{H\bar{y}} \sum_{h=1}^{H} y_h \omega_h - 1 \tag{13}$$

where y_h represents an outcome variable of interest for household h; $\bar{y}$ is the mean value of the outcome variable; and ω_h is the fractional rank of households along the income distribution, where 1 is at the bottom and H is the top of this distribution.

The concentration index must be normalised to analyse a binary variable on the -1 to +1 scale. This is the case for many of the x_k predictor variables outlined in Section 5.2.2. Two different approaches, proposed by Wagstaff and Erreygers, are available to carry this out. (for discussion, see Erreygers and Ourti, 2011). We follow Walsh and Cullinan (2015) and employ the Wagstaff normalisation (CI_{Wag}) which takes the following form:

$$CI_{Wag} = \frac{CI}{1 - \bar{y}}. \tag{14}$$

5.3 Empirical application results

Table 7 presents the predictive power of the CMC procedure. Each of these findings will now be discussed in turn. First, relative bias and RMSE are low and within acceptable bounds. We find that there is a relative bias of -0.047 for the MPI-LA indicator. This compares favourably to measures of bias found in the literature. Das and Chambers (2017), for instance, find a relative bias of 0.33 for the headcount poverty ratio in a comparable SAE method, while Pokhriyal and Jacques (2017) find a RMSE of 0.8 in their Senegal application. We see a correlation coefficient of 0.99 for both poverty headcount and MPI-LA indicators. Correlation is slightly lower, but also still within acceptable bounds, for poverty depth. At 0.926, this suggests that there is a residual degree of unobserved spatial heterogeneity in the intra-household distribution of resources that leads to a slightly lesser correlation coefficient than that for poverty headcount and MPI-LA results.

We consider the predictive power of the estimate as the proportion of true municipio-level MPI-LA poverty rates that fall within the 95% and 99% confidence intervals. MPI poverty depth performs marginally better than MPI poverty headcount in this regard. Nevertheless, performance by all metrics is strong, with the actual poverty metric falls within the confidence interval at a rate above 91% when measured at the 95% confidence interval and in excess of 97% when measured at the 99% confidence interval. This compares favourably with predictive power found in the literature. For comparison, Pokhriyal and Jacques (2017) found a correlation coefficients in the range of 0.84-0.91 for a small area estimation procedure that uses machine learning methods and is applied to a Senegalese application.

Indeed, the implications that this error may have for policy decision-making is shown to be minimal in Figure 4, where the difference between simulated and actual poverty rates is almost indistinguishable. It is clear to see that the CMC procedure captures the extremities quite well, effectively capturing the heterogeneity in social outcome. Where prediction inaccuracies occur, they are infrequent and small in magnitude. Figure 2 shows that a strong prediction is observed for the majority of small areas. As such, the discriminate sampling employed by the CMC procedure is effective in capturing the between-area heterogeneity associated with Mexican regions, as outlined by Tarozzi and Deaton (2009).

To give futher insight into the distribution of estimator precision, Figure 5 provides density distributions of the distance between actual and observed MPI-LA poverty rates by small area. We see that these errors are normally distributed with the majority of error within two standard deviations. There is a slight bias towards underestimation as predicted by Table 7, suggesting that simulated findings should be interpreted as a lower bound; expected poverty rates, on average, are at least the value simulated, with an average bias of -0.047 for the MPI-LA estimator.

There are a few points to note in relation to simulation performance. First, there is a greater error at the upper limit of reported MPI values when one examines Figure 2, suggesting that point estimates for the CMC procedure are marginally better for regions with lower poverty rates. However, results stay within acceptable bounds. Figure 3 gives some insight into the drivers of this variation. Deprivation depth for the Mexican sample exists on a pretty small scale, with the margin of error pretty homogeneous within that range, with extremely high values showing a wider margin of error. Deprivation headcount shows a progressive widening of error. Therefore, one may deduce that the wider margin of error for regions with extremely high rates of deprivation is driven by both errors in the estimation of depth and headcount poverty rates, with the error for regions with medium to medium-high rates of deprivation driven primarily by error in headcount poverty rates.

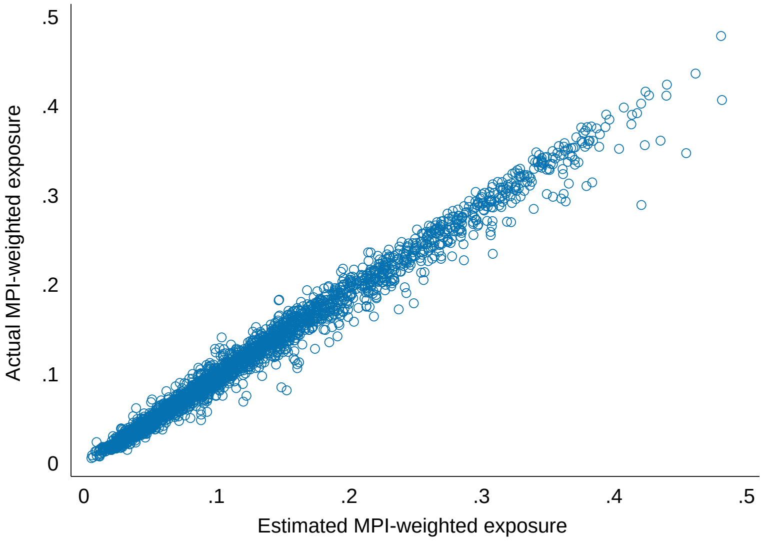
Figure 6 uses the concentration index methodology to project the correlation between MPI-quantified vulnerability at the municipio-level and key sociocononomic variables. Sanitation and food are chosen for illustrative purposes. As expected, we see a pro-vulnerable distribution for both poverty components. There is a much greater pro-vulnerability gradient with respect to food, suggesting that this likely of greater importance in alleviating municipio-level vulnerability than sanitation. Policy practitioners may use this and similar techniques to identify the most pertinent drivers of small area vulnerability, identifying the factors and their locations to target effective intervention. Should one wish to identify what regions to target, municipio level deprivation rates for drivers may be mapped, as has been shown in the Appendix. Similar to Figure 8, we can see that there is a strong positive correlation between true and predicted rates of vulnerability at the individual indicator level.

Table 7: Predictive power of CMC procedure

	Poverty indicator		
	MPI Poverty Headcount	MPI Poverty Depth	MPI-LA
Simulation error			
Relative Bias	-0.042	-0.003	-0.047
RMSE	0.031	0.014	0.014
Correlation - actual vs. simulated			
Pearson correlation	0.991	0.926	0.992
Spearman correlation	0.990	0.916	0.992
Estimate precision			
95% Confidence Interval	0.934	0.972	0.914
99% Confidence Interval	0.985	0.995	0.974

Note: Results derived from 100 simulation iterations. Estimate precision is calculated as the proportion of simulated municipios for which the real value lies within the simulated 95% and 99% confidence interval, calculated as sample mean + 2 standard deviations and the sample mean + 3 standard deviations, respectively. The calculated RMSE is the square root of the mean value of the 100 squared deviations of estimated and actual poverty. The calculated relative bias is the the mean value of the 100 proportional deviations of estimated and actual poverty.

Figure 2: Multidimensional Poverty Index (MPI-LA): Simulated vs. Actual

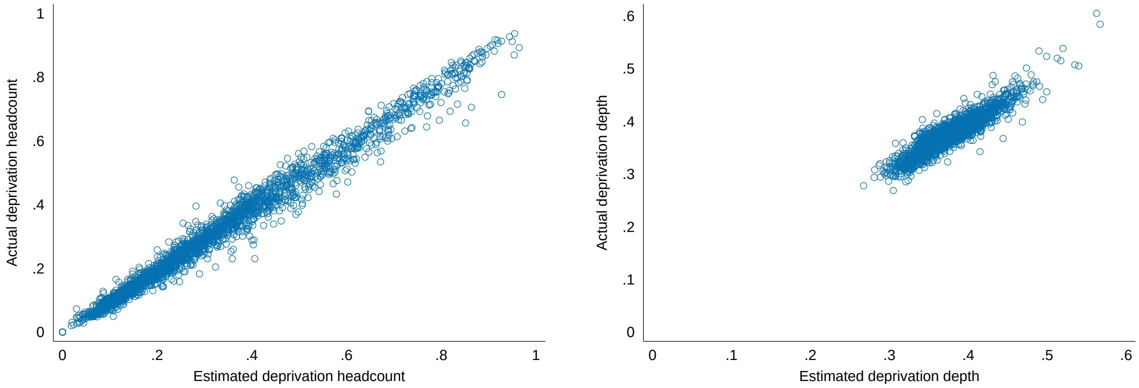


Predictive power of the CMC method. Figure shows comparison of actual and predicted MPI values for all municipios.

Figure 3: Deprivation depth and deprivation headcount: simulated vs. actual

(a) Deprivation headcount

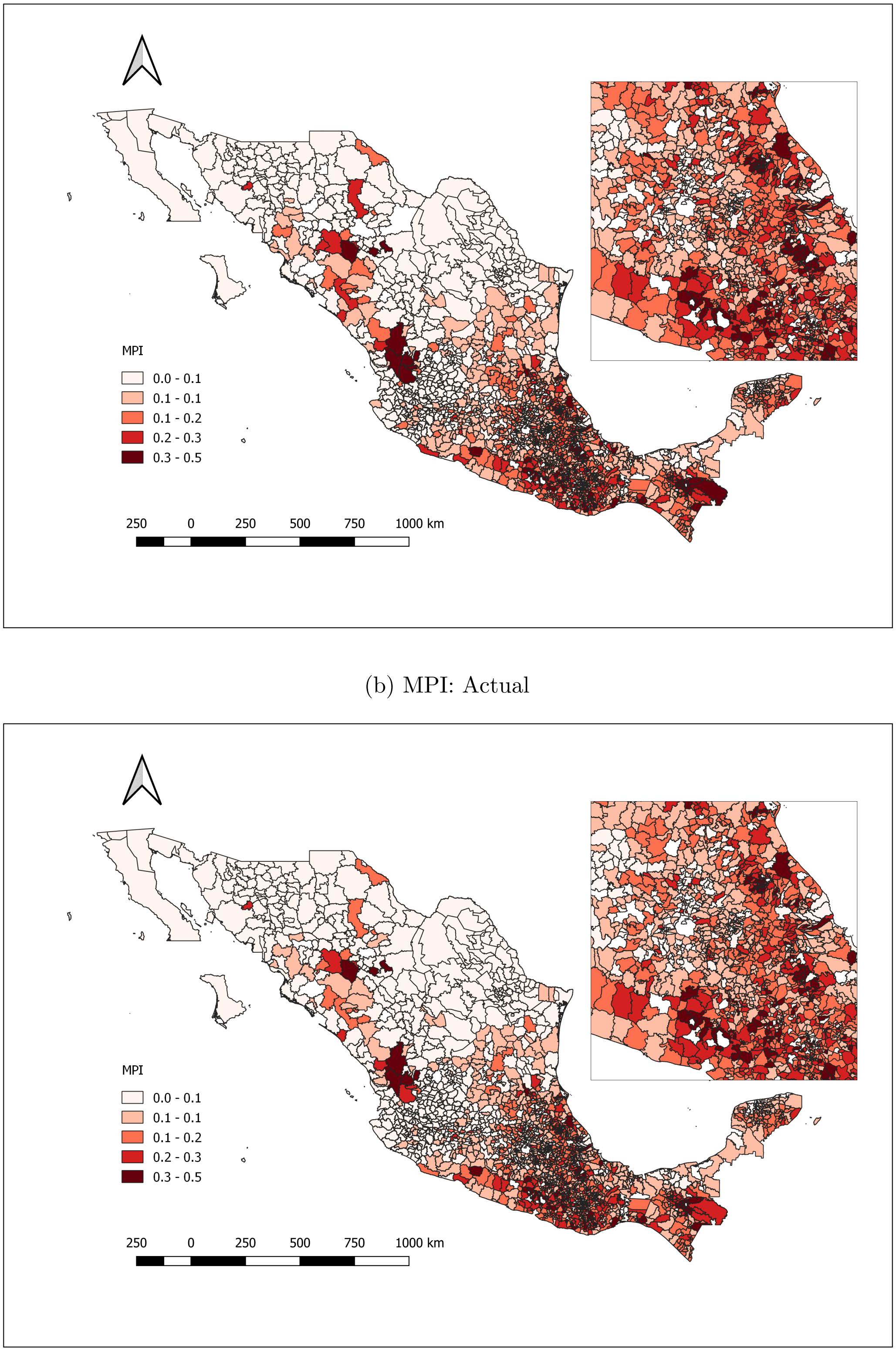
(b) Deprivation depth



Predictive power of the CMC method. Figure shows comparison of actual and predicted Deprivation headcount and deprivation depth for all municipios.

Figure 4: MPI by municipio

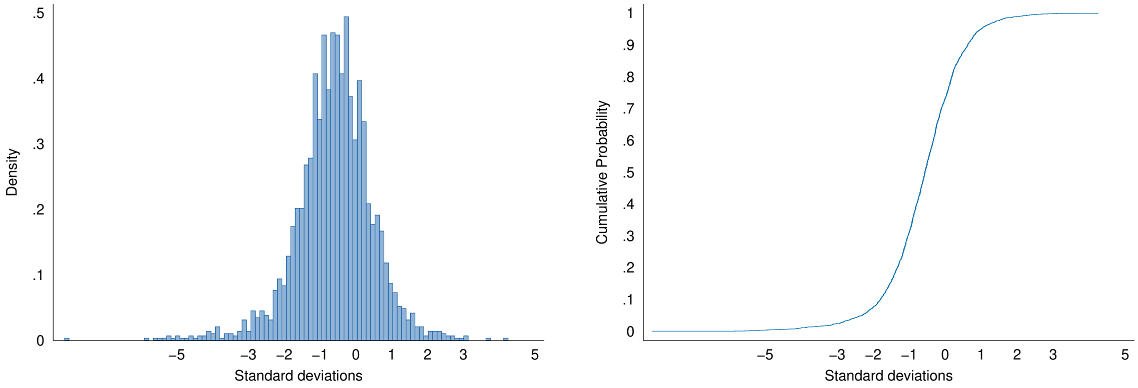
(a) MPI: Simulated



(b) MPI: Actual

Note: Simulated and actual MPI displayed by municipio. Municipios of southern region enlarged. Categorisation according to natural breaks (jenks) in the data.

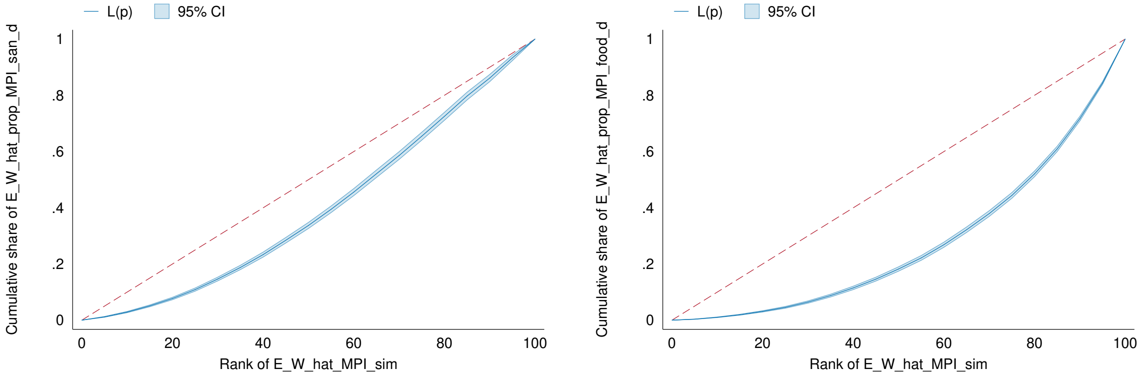
Figure 5: Estimator precision: distance between actual and observed small area poverty rate


(a) Precision histogram

(b) Cumulative distribution

Note: Confidence interval coverage calculated as the number of municipios for which the true MPI-LA poverty value lies within the stated distance from the simulated mean, calculated in terms of standard deviations.

Figure 6: Distribution of poverty indicators across deprivation spectrum


(a) Sanitation

(b) Food

Note: Concentration index methodology used to quantify social gradient of climate impact incidence.

6 Conclusion

Small area poverty estimates are important for social and economic policy however the required data are often unavailable. A number of estimation procedures have emerged since the original poverty mapping methodology proposed by Elbers et al. (2003), with each procedure requiring careful consideration of data distribution and modelling parameters (Molina and Rao, 2010; Pfeffermann, 2013; Rao, 2003) to ensure unbiased inference. This paper has introduced a Conditional Monte Carlo (CMC) procedure that provides robust estimates of small area poverty and is subject to fewer restrictive assumptions. We present a theoretical derivation, a numerical validation of concept and an empirical application to a Mexican case study. We demonstrate that the CMC method estimates small area poverty rates with precision.

The CMC method expands upon similar survey reweighting methods by providing an estimate of precision, while the microdata-based estimation procedure widens the scope of insight relative to common small area estimation procedures. Between-area heterogeneity in the relationship between outcome and predictor variables is accounted for through a discriminate sampling procedure. Using a Mexican case study, this paper demonstrates the simulation of small area poverty estimates with high degrees of precision. While Tarozzi and Deaton (2009) and Tarozzi (2011) show that between area heterogeneity may create estimate bias when traditional SAE methods are employed, our results suggest that discriminate sampling has overcome this limitation.

The findings of this paper have important policy implications. Estimates to calculate welfare at the small area are widely used by policy decision-makers, with institutions such as the World Bank investing heavily in the development of both methods and analyses to aid the targeting of poverty relief. However, there is considerable uncertainty in the literature as to the most appropriate method to be applied to a given context. The CMC method introduced in this paper presents greater transparency and simplicity in estimation, potentially widening the applicability of robust small area estimation.

References

Aaberge, R., E. Peluso, and H. Sigstad (2019, sep). The dual approach for measuring multidimensional deprivation: Theory and empirical evidence. *Journal of Public Economics 177*, 104036.

Alkire, S., P. Conceição, A. Barham, C. Calderón, A. Conconi, J. Dirksen, F. C. Espinal, M. Evans, J. Hall, A. Jahic, U. Kanagaratnam, M. Kivilo, M. Kovacevic, F. Kovesdi, C. Mitchell, R. Nogales, C. Oldiges, A. Ortubia, M. Pinilla-Roncancio, C. Rivera, M. E. Santos, S. Scharlin-Pettee, S. Seth, A. Vaz, F. Vollmer, and C. Walkey (2019). *The Global Multidimensional Poverty Index (MPI) 2019: Illuminating Inequalities.*

Alkire, S. and J. Foster (2011, aug). Counting and multidimensional poverty measurement. *Journal of Public Economics 95*(7-8), 476–487.

Atkinson, A. B. (2003). Multidimensional deprivation: contrasting social welfare and counting approaches. *The Journal of Economic Inequality 1*, 51–65.

Ballas, D., G. Clarke, and J. Dewhurst (2006). Modelling the socio-economic impacts of major job loss or gain at the local level: a spatial microsimulation framework. *Spatial Economic Analysis 1*(1), 127–146.

Ballas, D. and G. P. Clarke (2001). Modelling the local impacts of national social policies: a spatial microsimulation approach. *Environment and Planning C: Government and Policy 19*(4), 587–606.

Ballas, D., G. P. Clarke, and E. Wiemers (2005). Building a dynamic spatial microsimulation model for ireland. *Population, Space and Place 11*(3), 157–172.

Bedi, T., A. Coudouel, and K. Simler (Eds.) (2007, sep). *More than a Pretty Picture: Using Poverty Maps to Design Better Policies and Interventions.* World Bank Publications.

Bourguignon, F. and A. Spadaro (2006, jan). Microsimulation as a tool for evaluating redistribution policies. *The Journal of Economic Inequality 4*(1), 77–106.

Burchi, F., D. Malerba, N. Rippin, and C. E. Montenegro (2019). Comparing global trends in multidimensional and income poverty and assessing horizontal inequalities. Technical report, Discussion Paper.

Cardaci, A. (2018, may). Inequality, household debt and financial instability: An agent-based perspective. *Journal of Economic Behavior & Organization 149*, 434–458.

Chambers, R. and N. Tzavidis (2006). M-quantile models for small area estimation. *Biometrika 93*(2), 255–268.

Curtis, B. (2018). Multidimensional poverty measurements. In *Understanding Global Poverty*, pp. 47–71. Routledge.

Das, S. and R. Chambers (2017, aug). 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 (Statistics in Society) 180*(4), 1137–1161.

Devarajan, S. (2013). Africa's statistical tragedy. *Review of Income and Wealth 59*(S1), S9–S15.

Diallo, M. S. (2014). *Small Area Estimation under Skew-Normal Nested Error Models.* Ph. D. thesis, Ottawa-Carleton Institute for Mathematics and Statistics, Carleton University, Canada.

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

Elbers, C. and R. van der Weide (2014, July). Estimation of normal mixtures in a nested error model with an application to small area estimation of poverty and inequality. Policy Research Working Paper Series 6962, The World Bank.

Erreygers, G. and T. V. Ourti (2011, jul). Measuring socioeconomic inequality in health, health care and health financing by means of rank-dependent indices: A recipe for good practice. *Journal of Health Economics 30*(4), 685–694.

Farmer, J. D. and D. Foley (2009, aug). The economy needs agent-based modelling. *Nature 460*(7256), 685–686.

Farrell, N., K. Morrissey, and C. O'Donoghue (2013). *Creating a Spatial Microsimulation Model of the Irish Local Economy*, pp. 105–125. Dordrecht: Springer Netherlands.

Ghosh, M., J. Rao, et al. (1994). Small area estimation: an appraisal. *Statistical science 9*(1), 55–76.

Hall, O., F. Dompae, I. Wahab, and F. M. Dzanku (2023). A review of machine learning and satellite imagery for poverty prediction: Implications for development research and applications. *Journal of International Development n/a*(n/a).

Hallegatte, S. and J. Rozenberg (2017). Climate change through a poverty lens. *Nature Climate Change 7*(4), 250.

Happe, K., A. Balmann, K. Kellermann, and C. Sahrbacher (2008, aug). Does structure matter? the impact of switching the agricultural policy regime on farm structures. *Journal of Economic Behavior & Organization 67*(2), 431–444.

Harding, A. (2017). *New frontiers in microsimulation modelling*. Routledge.

Hentschel, J., J. O. Lanjouw, P. Lanjouw, and J. Poggi (2000). Combining census and survey data to trace the spatial dimensions of poverty: A case study of ecuador. *The World Bank Economic Review 14*(1), 147–165.

Lovelace, R. and D. Ballas (2013, sep). 'truncate, replicate, sample': A method for creating integer weights for spatial microsimulation. *Computers, Environment and Urban Systems 41*, 1–11.

Minot, N. (2000). Generating disaggregated poverty maps: An application to vietnam. *World development 28*(2), 319–331.

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

Morrissey, K., C. O'donoghue, and N. Farrell (2014). The local impact of the marine sector in ireland: a spatial microsimulation analysis. *Spatial Economic Analysis 9*(1), 31–50.

Naraya, D., R. Patel, K. Schafft, A. Rademacher, and S. Koch-Schulte (2000, mar). *Can Anyone Hear Us?* The World Bank.

Narayan-Parker, D. and R. Patel (2000). *Voices of the poor: Can anyone hear us?*, Volume 1. World Bank Publications.

Norman, P. (1999). Putting iterative proportional fitting on the researcher's desk. *School of Geography, University of Leeds Working Paper 99*(03).

O'Donoghue, C., N. Farell, K. Morrissey, J. Lennon, D. Ballas, G. Clarke, and S. Hynes (2013). *The SMILE Model: Construction and Calibration*, pp. 55–86. Berlin, Heidelberg: Springer Berlin Heidelberg.

O'Donoghue, C., K. Morrissey, and J. Lennon (2014). Spatial microsimulation modelling: a review of applications and methodological choices.

OPHI (2013). Measuring multidimensional poverty: Insights from around the world. *OPHI Briefing Paper*.

Permanyer, I. (2014, jul). Assessing individuals' deprivation in a multidimensional framework. *Journal of Development Economics 109*, 1–16.

Pfeffermann, D. (2013, feb). New important developments in small area estimation. *Statistical Science 28*(1), 40–68.

Pham, A. T. Q., P. Mukhopadhaya, and H. Vu (2020). Targeting administrative regions for multidimensional poverty alleviation: A study on vietnam. *Social Indicators Research 150*(1), 143–189.

Pokhriyal, N. and D. C. Jacques (2017). Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences 114*(46), E9783–E9792.

Pratesi, M., L. Quattrociocchi, G. Bertarelli, A. Gemignani, and C. Giusti (2021). Spatial distribution of multidimensional educational poverty in italy using small area estimation. *Social Indicators Research 156*, 563–586.

Rahman, A. and A. Harding (2016). *Small area estimation and microsimulation modeling*. Chapman and Hall/CRC.

Rahman, A., A. Harding, R. Tanton, S. Liu, et al. (2010). Methodological issues in spatial microsimulation modelling for small area estimation. *International Journal of Microsimulation 3*(2), 3–22.

Rao, J. N. K. (2003, Jan). *Small Area Estimation*. John Wiley & Sons, Inc.

Rao, N. D., B. J. van Ruijven, K. Riahi, and V. Bosetti (2017, nov). Improving poverty and inequality modelling in climate research. *Nature Climate Change 7*(12), 857–862.

Reddy, S. G. and T. Pogge (2002). How not to count the poor.

Salvati, N., N. Tzavidis, M. Pratesi, and R. Chambers (2010, dec). Small area estimation via m-quantile geographically weighted regression. *TEST 21*(1), 1–28.

Salvati, N., N. Tzavidis, M. Pratesi, and R. Chambers (2012). Small area estimation via m-quantile geographically weighted regression. *Test 21*(1), 1–28.

Santos, M. E. and P. Villatoro (2018). A multidimensional poverty index for latin america. *Review of Income and Wealth 64*(1), 52–82.

Sen, A. (1983). Poor, relatively speaking. *Oxford Economic Papers 35*(2), 153–169.

Sen, A. (1999). *Development as Freedom*. Oxford University Press.

Sen, A. K. (1992). *Inequality reexamined*. Oxford University Press.

Sen, A. K. (2009). *The idea of justice*. Harvard University Press.

Simler, K. (2016). *Pinpointing Poverty in Europe : New Evidence for Policy Making.* Washington, DC.: World Bank.

Smith, D. M., G. P. Clarke, and K. Harland (2009). Improving the synthetic data generation process in spatial microsimulation models. *Environment and Planning A 41*(5), 1251–1268.

Stiglitz, J., A. K. Sen, and J.-P. Fitoussi (2009, December). The measurement of economic performance and social progress revisited: Reflections and Overview. (2009-33).

Tanton, R. and K. Edwards (2012). *Spatial microsimulation: A reference guide for users*, Volume 6. Springer Science & Business Media.

Tarozzi, A. (2011, jul). Can census data alone signal heterogeneity in the estimation of poverty maps? *Journal of Development Economics 95*(2), 170–185.

Tarozzi, A. and A. Deaton (2009, nov). Using census and survey data to estimate poverty and inequality for small areas. *Review of Economics and Statistics 91*(4), 773–792.

UN (2015). United nations sustainable development goals. *https://www.un.org/sustainabledevelopment/sustainable-development-goals/*.

UNDP (2013). *A million voices: The world we want. A sustainable future with dignity for all.* New York: UNDP.

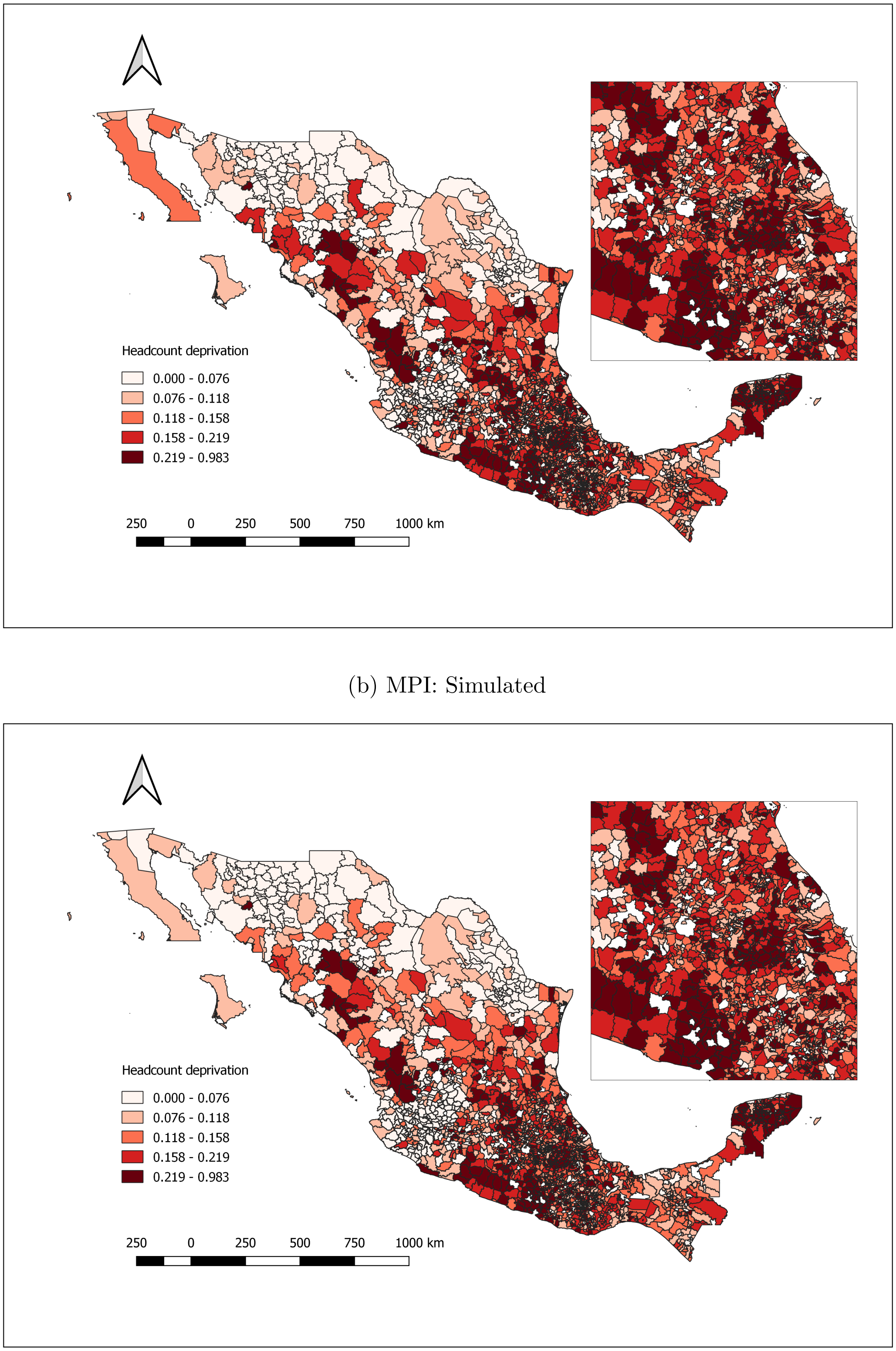
Walsh, B. and J. Cullinan (2015, jan). Decomposing socioeconomic inequalities in childhood obesity: Evidence from ireland. *Economics & Human Biology 16*, 60–72.

Whitworth, A., E. Carter, D. Ballas, and G. Moon (2017, may). Estimating uncertainty in spatial microsimulation approaches to small area estimation: A new approach to solving an old problem. *Computers, Environment and Urban Systems 63*, 50–57.

Yeh, C., A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications 11*(1), 2583.

7 Appendix: Spatial profile of indicator variables

Figure 7: Sanitation: actual vs simulated headcount deprivation

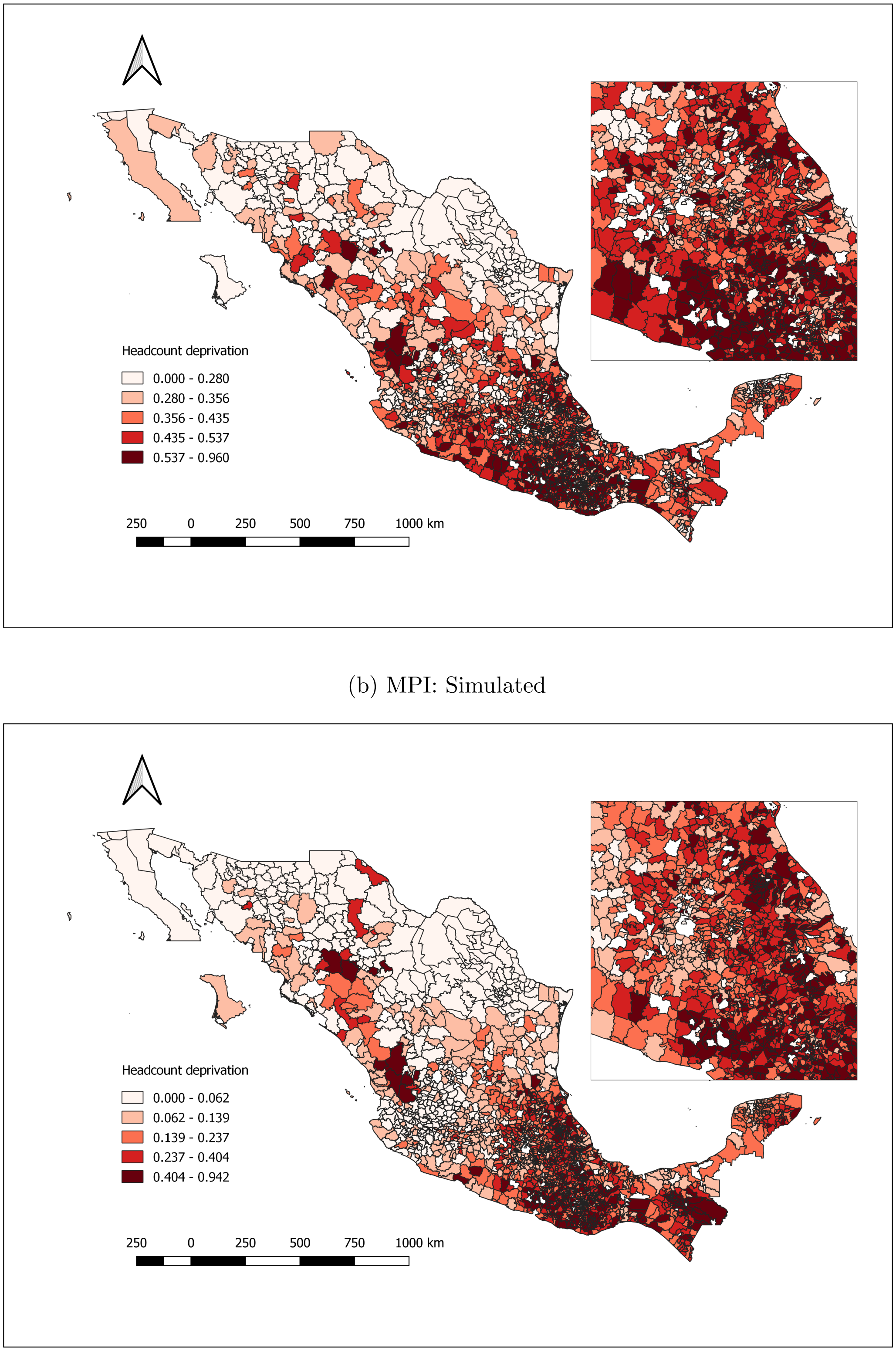


(a) MPI: Actual

(b) MPI: Simulated

Note: Simulated and actual MPI displayed by municipio. Municipios of southern region enlarged. Categorisation according to natural breaks (jenks) in the data.

Figure 8: Insufficient food: actual vs simulated headcount deprivation



(a) MPI: Actual

(b) MPI: Simulated

Note: Simulated and actual MPI displayed by municipio. Municipios of southern region enlarged. Categorisation according to natural breaks (jenks) in the data.