USING CENSUS AND SURVEY DATA TO ESTIMATE POVERTY AND INEQUALITY FOR SMALL AREAS

Alessandro Tarozzi and Angus Deaton*

Abstract—Recent years have seen widespread use of small-area maps based on census data enriched by relationships estimated from household surveys that predict variables, such as income, not covered by the census. The purpose is to obtain putatively precise estimates of poverty and inequality for small areas for which no or few observations are available in the survey. We argue that to usefully match survey and census data in this way requires a degree of spatial homogeneity for which the method provides no basis and which is unlikely to be satisfied in practice. We document the potential empirical relevance of such concerns using data from the 2000 census of Mexico.

I. Introduction

HOUSEHOLD surveys collect information on incomes, expenditures, and demographics and are regularly used to generate population statistics, such as mean incomes, poverty head count ratios, or rates of malnutrition. Such surveys are now widely available around the world. For example, in its latest estimates of the global poverty counts, the World Bank used 454 income and expenditure surveys from 97 developing countries (Chen & Ravallion, 2004). Some of these surveys support subnational estimates, for example, for states or provinces. But few surveys are large enough to support estimates for small areas such as districts, counties, school districts, or electoral constituencies. In the United States, where the decadal census collects good income information for 5% of the population, there is a substantial literature, including two National Research Council reports, on obtaining mean income and poverty estimates for counties and school districts in the intercensal years, estimates that are required for the apportionment of federal funds (National Research Council, 1980; Grosh & Rao, 1994; Citro & Kalton, 2000).

In most developing countries, censuses do not collect income or expenditure information, so small-area poverty estimates are typically not available even for census years. To fill this gap, the World Bank has recently invested in a methodology for generating small-area poverty and inequality statistics, in which an imputation rule, estimated from a household survey, is used to calculate small-area estimates from census data. The methodology, developed by Elbers, Lanjouw, and Lanjouw (2003, henceforth ELL), has been applied, with some local variation, to a number of countries, including Albania, Azerbaijan, Brazil, Bulgaria, Cambodia, China, Ecuador, Guatemala, Indonesia, Kenya, Madagascar,

Mexico, Morocco, South Africa, Tanzania, and Uganda.[1] In many cases, and even when the area is as small as a few thousand people, the estimates come with high reported precision; for example, the Kenyan poverty map reports poverty rates for areas with as few as 10,000 people with relative standard errors of a quarter and of around 10% for areas with 100,000 people. In some cases, such as Kenya, the provision of poverty maps has become part of the regular statistical service.[2] In others, hundreds of millions of dollars have been distributed based on the estimates. And the computed poverty and inequality estimates have been used in other studies, for example, of project provision and political economy, the effects of inequality on crime, whether inequality is higher among the poor, and child malnutrition.

In spite of the widespread application and growing popularity of poverty mapping, there has been little formal investigation of its properties. The original paper by ELL describes their procedure, but does not provide a characterization of the general properties on which the imputation is based or a consideration of the likelihood or consequences of assumption failure.

In this paper, we provide a set of conditional independence or area homogeneity assumptions that are required for the poverty mapping to provide useful estimates for small areas. These assumptions, which are closely related to the ignorability or unconfoundedness assumptions familiar from the statistical and econometric literature on program evaluation, require that at least some aspects of the conditional distribution of income be the same in the small area as in the larger area that is used to calibrate the imputation rule. We argue that the area homogeneity assumptions are likely to fail in practice and that local labor markets, local rental markets, and local environmental differences are likely to generate heterogeneity that violates the assumptions of both the ELL estimator and a simpler version that we propose below. More generally, we note that the imputation formulas are projections of expenditure, income, or poverty on a subset of whatever variables happen to be common to the census and the survey, supplemented by local averages from the census, and are not well-founded structural relationships, so their coefficients will generally be functions of any local variables that are not explicitly included.

We consider some obvious special cases of heterogeneity, where either the intercept or the slopes of the projection

[1] For a comprehensive description of the methodology that the World Bank used, as well as for reference to the numerous applications, see http://go.worldbank.org/9CYUFEUQ30.

[2] See http://go.worldbank.org/K5P2DUGH20.

vary randomly across areas, and discuss the consequences for estimators. We focus on mean squared error (MSE) and coverage probabilities rather than means, since in many cases of interest, including the example above, the estimates are not consistent. While both ELL and our own estimators produce precise estimates of welfare measures in some cases, we also show that even a small amount of heterogeneity may lead to misleading inference.

We provide calculations from the Mexican Census of 2000, which we use to construct random synthetic household surveys that are used to calculate imputation rules for poverty. Because the Mexican census contains income information, these can be checked against the poverty rates for small areas calculated from the census extract. While the poverty mapping technique is certainly informative in this case, the coverage probabilities are often far from the nominal ones, so that for a substantial fraction of the areas we consider, nominal standard errors based on homogeneity provide misleading indications of precision.

The rest of the paper is organized as follows. The next section introduces the notation, formally describes the problem, and discusses the assumptions that justify merging census and survey data. Section III describes the consequences of unobserved heterogeneity across areas. Section IV describes estimation and proposes a simple estimator that is appropriate under the same assumptions as the estimator in Elbers et al. (2003). Section V describes a series of Monte Carlo simulations designed to explore the consequences of unobserved heterogeneity. Section VI describes the validation exercise with data from the 2000 Mexican Census. Section VII concludes.

II.   Statistical Background

The object of interest is a welfare measure W defined for a small area A, where $A \subset R$ denotes a small area included in a larger region R. For instance, A may be a town and R a district, or A may be a district while R is a state. In a typical census, each small area will be further divided into a number of smaller units or clusters, usually referred to as census tracts or enumeration areas (EAs), typically containing around 100 households. In this paper, we use the term *cluster* throughout, and we treat *cluster* and *EA* as synonymous. In most cases, W is a poverty or inequality index defined as a function of the distribution over individuals of a variable y, which usually measures income or expenditure (*expenditure* hereafter). However, W may also be a function of the distribution of other variables, such as wages, schooling, or occupation or health indicators. In the frequent case where data on y are collected at the household level, we assign to each individual within a household the same per capita measure y.

Most poverty measures are identified by a simple population moment condition such as

$$E[s_h g(y_h; W_0)|h \in H(A)] = 0, \qquad (1)$$

where s_h represents the size of household h, W_0 is the true value of the parameter to be estimated, and $H(A)$ denotes the set of households in area A. For instance, if W_0 represents a Foster-Greer-Thorbecke (FGT) poverty index and z is a fixed poverty line, then

$$g(y_h; W_0) = 1(y_h < z)\left(1 - \frac{y_h}{z}\right)^{\alpha} - W_0, \qquad (2)$$

where $\alpha \geq 0$ is a known parameter and $1(E)$ is an indicator equal to 1 when event E is true. When $\alpha = 0$, the index becomes the head count poverty ratio, while $\alpha = 1$ characterizes the poverty gap ratio. A larger parameter α indicates that large poverty gaps $(1 - y/z)$ are given a larger weight in the calculation, so that the poverty index becomes more sensitive to the distribution of y among the poor. Most inequality measures can be written as continuous functions of expected values, each identified by a moment condition. For instance, the variance of the logarithms can be written as $E[(\ln y)^2] - [E(\ln y)]^2$. The Theil inequality index is defined as $E[y \ln y]/E(y) - \ln(E(y))$, and the Atkinson inequality index is

$$W_0 = 1 - \frac{E(y^{1-\epsilon})^{(1/1-\epsilon)}}{E(y)}.$$

The Gini coefficient, using a formula described in Dorfman (1979), can also be written in terms of elements identified by a moment condition as

$$W_0 = 1 - \frac{\int_0^{\infty}(1 - E[1(y \leq z)])^2 dz}{E(y)}.$$

There are two data sources. The first is a household survey of region R that includes data on y as well as on a set of correlates X. We assume that the sample size allows the estimation of aspects of the distribution of y in region R with acceptable precision where what is acceptable will depend on specific circumstances. For instance, the precision of the resulting welfare estimates for region R could be deemed acceptable if it allows sufficient power in tests that compare welfare estimates for region R with estimates from other regions, or from the same region but in a different period. The second data source is a census of the whole population of households $h \in H(A)$. The census will usually include information from a larger area (such as the whole region R), but for our purposes, only data from the small area $A \subset R$ are relevant. We assume that the census does not include information on expenditure y, but it does record information on the correlates X. Note that the choice of correlates, while influenced by theory, is ultimately constrained by the overlap between census and survey, each designed with other purposes in mind.

If y is recorded for a sample of households in area A, the welfare estimate W_0 can be estimated using a sample analog

of the corresponding moment condition. As an example, the FGT poverty index can be estimated as

$$\hat{W}_0 = \frac{1}{\sum_{h \in H_n(A)} s_h} \sum_{h \in H_n(A)} s_h 1(y_h < z)\left(1 - \frac{y_h}{z}\right)^{\alpha}, \quad (3)$$

where $H_n(A)$ denotes the set of households from area A included in the survey sample. Under fairly general regularity conditions, such an estimator is consistent and asymptotically normal. However, the corresponding standard errors will be large if the number of observations is small, a common circumstance if the area A is only a small subset of the larger area covered by the survey that collects information on y. The survey may indeed include no households at all from certain areas. Sample size would be more than adequate in a complete census of the small area, which will typically include several thousand households. Censuses, though, rarely include reliable information on income or expenditure. However, a census will record a list of variables X, such as occupation, schooling, housing characteristics, or availability of amenities at the local level, which are also recorded in household surveys and can be used as predictors for y. If the survey also includes detailed geographical identifiers, one can also calculate averages of household-level variables calculated for small locations (e.g., a village) and attach these variables to the survey data as additional predictors of expenditure (Elbers et al., 2003). Under certain conditions, one can then merge information from both data sets to improve the precision of the estimates of W_0 for a small area A. Consider the following assumptions:

Assumption 1 (MP): Measurement of Predictors: Let X_h denote the value of the correlates for household h as observed if h is included in the survey sample, and let $\tilde{X}_h$ denote the corresponding measurement in the census. Then $X_h = \tilde{X}_h$ for all h.

Assumption 2 (AH): Area Homogeneity (or Conditional Independence):

$$f(y_h|X_h, h \in H(A)) = f(y_h|X_h, h \in H(R)). \quad (4)$$

Assumption MP is clearly necessary if the correlates have to be used to bridge census and household data. The validity of this assumption should not be taken for granted. For instance, the two data sources may use a different definition of *household,* or they may use different (possibly nonnested) coding schemes for schooling, industry, or occupation of household members. Different reports may also arise from less obvious reasons, even if census and household survey use the same wording to record all variables included in X. For instance, reporting errors may differ due to differences in questionnaire length or interviewer training.[3] In the rest

of the paper, we will maintain the validity of MP, but the caveats just described should be kept in mind. We also assume that the list of correlates that are measured consistently in the two surveys also includes household size, but none of our results rely crucially on this assumption.

Assumption AH requires that the conditional distribution of y given X in the small area A is the same as in the larger region R. Conditional independence assumptions such as AH have been used extensively in statistics and econometrics. Following the seminal work by Rubin (1974) and Rosenbaum and Rubin (1983), the program evaluation literature has made frequent use of the assumption (sometimes referred to as unconfoundedness or ignorability) that treatment status is independent of potential outcomes, conditionally on observed covariates (see, e.g., the references surveyed in Heckman, LaLonde, & Smith, 1999, and Imbens, 2004). In the estimation of models with missing data, several authors have used the identifying assumption that the probability of having a complete observation conditional on a set of auxiliary variables is constant (see, e.g., Rubin, 1976; Little & Rubin, 2002; Wooldridge, 2002b). Analogous assumptions can be found in the estimation of nonlinear models with nonclassical measurement errors in the presence of validation data. In this case, the requirement is that the distribution of the mismeasured variables conditional on a set of proxies is the same in the main and in the auxiliary sample (see, e.g., Lee & Sepanski, 1995; Chen, Hong, & Tamer, 2005; Chen, Hong, and Tarozzi, 2008).

In the estimation of small-area statistics, assumption AH is demanding due to the many possible sources of heterogeneity in the relationship between the predictors and y across different areas. For example, X may include schooling or occupation variables, but the conditional relation between such factors and expenditure is driven by local rates of return, which are typically unobserved and unlikely to be identical across different geographical areas. The inclusion of physical assets, or proxies for physical assets, such as indicators of durable ownership, may capture some of the variation in the rates of return. However, such indicators are subject to similar concerns because the rate of return to assets may vary across areas. Differences in tastes, relative prices, or the environment across areas will also lead to the failure of AH; the implications of bicycle or television ownership for the poverty of a household must depend on whether the area is suitable for riding a bicycle or whether the village has an electricity supply or television signal. It should also be noted that the conditional distribution will generally change over time, so caution should be exercised when survey and census data have not been collected during the same period. This is a common circumstance, because while censuses are usually completed only once every decade, household expenditure surveys are often completed at shorter intervals. More generally, the coefficients of the projection of y on X, including the constant

---

[e] See Deaton and Grosh (2000) for a brief overview of the difficulties related to reporting bias in household surveys.

term, will be a function of omitted variables; if these are not constant across localities, area homogeneity will fail.

The area homogeneity assumption AH requires, for instance, that the probability of being poor given X in the small area A is the same as in the larger region R. If assumptions MP and AH hold, the welfare estimate of interest is also identified by the following modified moment condition:

$$\int E[s_h g(y_h; W_0)|X, h \in H(R)]dF(X|h \in H(A)) = 0, \quad (5)$$

where $dF(\cdot)$ represents the distribution of the correlates in the small area.[4] In Appendix A we show that (5) can be obtained from assumptions MP and AH from a simple manipulation of the moment condition (1). If we replace the modified moment condition (5) by its sample analog, we have a basis for estimating the welfare measure. As the sample size within each area becomes large, the sample analog will converge to (1) and give a consistent estimate of the welfare measure. In practice, with a finite number of households in each area, consistency will not guarantee estimator precision, but it provides a basis from which we can examine performance in terms of MSE.

III.   Consequences of Unobserved Heterogeneity

In this section, we maintain the validity of MP while we discuss consequences of the presence of unobserved heterogeneity, which invalidates AH. Virtually all household expenditure surveys adopt a complex survey design, so that EAs such as villages and urban blocks are sampled first, and then households are sampled from each EA. As is well known, the resulting intracluster correlation among households drawn from the same EA can considerably increase the standard errors of the estimates (see, e.g., Kish, 1965; Cochrane, 1977). In what follows, the subscript a denotes a small area, c denotes a cluster or primary stage unit, and h denotes a household. Hence, for instance, y_{ach} indicates expenditure of household h, residing in cluster c, inside area a. Every cluster is assumed to be completely included in a unique small area. For illustrative purposes, we abstract from the distinction between household and individual-level observations.

To fix ideas and more clearly illustrate the concepts, assume temporarily that the relationship between y and the correlates X is described by a parametric linear model whose coefficients, apart from the constant term, are homogeneous across areas. This provides the simplest example of a (limited) failure of area homogeneity. In reality, heterogeneity in the slopes is also likely and, as documented in section VB, equally capable of leading to incorrect infer-

ence. Suppose then that the data-generating process (DGP) is described by the following model:

$$y_{ach} = \beta'X_{ach} + u_{ach} = \beta'X_{ach} + \eta_a + e_{ac} + \varepsilon_{ach}, \quad (6)$$

where $\mathrm{Cov}(\eta_a, e_{ac}) = 0$, $\mathrm{Cov}(\eta_a, \varepsilon_{ach}) = 0$, $\mathrm{Cov}(e_{ac}, \varepsilon_{ach}) = 0$, $\mathrm{Cov}(X_{ach}, u_{ach}) = 0$, $\mathrm{Cov}(\varepsilon_{ach}, \varepsilon_{ach'}) = 0 \,\forall a, c, h, h' \neq h$, $\mathrm{Cov}(e_{ac}, e_{ac'}) = 0 \,\forall a, c, c' \neq c$, $\mathrm{Cov}(\eta_a, \eta_{a'}) = 0 \,\forall a, a' \neq a$. All error components are uncorrelated with each other and with the correlates. We assume that model (6) holds for every cluster c in region R, so that it also holds for all clusters within the small area. Model (6) allows for the presence of a small-area fixed effect η_a, which violates area homogeneity, but it otherwise maintains homogeneity in the slopes β, which can be consistently estimated using either ordinary least squares or feasible generalized least squares on survey data from the larger region R. Note that assumption AH fails because in a specific small-area A,

$$E(y_{ach}|X_{ach}, h \in H(A)) = \beta'X_{ach} + \eta_a \neq \beta'X_{ach}$$
$$= E(y_{ach}|X_{ach}, h \in H(R)).$$

In this case, because of the violation of homogeneity through the presence of η_a, we cannot obtain consistent estimation of welfare estimates for small areas by merging census and survey data. Suppose that the object of interest is the simple poverty head count for a small-area A, that is, $W_A = P(y \leq z|a = A)$, where z denotes the poverty line. The head count in A is equal to $P(y \leq z|a = A) = P(e_{ac} + \varepsilon_{ach} \leq z - \beta'X_{ach} - \eta_a)$, but without knowing η_a, this quantity cannot be calculated even if both β and the distribution of $e_{ac} + \varepsilon_{ach}$ were known. In such a case, the use of household survey data from the larger region R will not allow the consistent estimation of the welfare estimate W_A.

The presence of this kind of heterogeneity makes the problem of estimating W_A similar to the problem of making forecasts in time series analysis. In time series forecasting, while parameters that relate the predicted variables to their predictors can—under appropriate conditions—be estimated consistently, the same cannot be said for the actual (future) value of the variables to be predicted. For this reason, inference on the predictions should be based on measures of mean squared forecast error (MSE). In our context, the presence of the area fixed effect η_a, which cannot be precisely estimated without a large sample of observations (y_{ach}, X_{ach}) from the small area a, implies that the MSE of $\hat{W}_A$ will also be affected by the presence of bias. The following section illustrates the point further and describes the consequences for MSE of ignoring the presence of a small-area fixed effect under a variety of DGPs.

A.   Consequences of Area Heterogeneity for Mean Squared Error

As in the previous section, we assume that region R is composed of a number of small areas labeled a, each

---

[4] The validity of (5) also requires that the support of X in A is a subset of the support of X in R, but this condition holds by construction, because the small area is a subset of the larger region R.

TABLE 1.—EFFECT OF INTERCLUSTER CORRELATION ON ROOT MSE

	(1)	(2)	(3)	(4)	(5)	(6)
	$\rho_a = 0.1$					
	$C = 10$		$C = 150$		$C = 500$	
	$\eta = \tau_{\eta,.75}$	$\eta = \tau_{\eta,.90}$	$\eta = \tau_{\eta,.75}$	$\eta = \tau_{\eta,.90}$	$\eta = \tau_{\eta,.75}$	$\eta = \tau_{\eta,.90}$
(a) $\rho_c = 0.2$	1.6	2.9	5.8	10.9	10.5	19.9
	$\rho_a = 0.01$					
	$C = 10$		$C = 150$		$C = 500$	
	$\eta = \tau_{\eta,.75}$	$\eta = \tau_{\eta,.90}$	$\eta = \tau_{\eta,.75}$	$\eta = \tau_{\eta,.90}$	$\eta = \tau_{\eta,.75}$	$\eta = \tau_{\eta,.90}$
(b) $\rho_c = 0.02$	1.5	2.5	4.9	9.1	8.8	16.6
(c) $\rho_c = 0.05$	1.3	1.9	3.5	6.5	6.2	11.8
(d) $\rho_c = 0.20$	1.1	1.3	2.1	3.6	3.4	6.4
	$\rho_a = 0.005$					
	$C = 10$		$C = 150$		$C = 500$	
	$\eta = \tau_{\eta,.75}$	$\eta = \tau_{\eta,.90}$	$\eta = \tau_{\eta,.75}$	$\eta = \tau_{\eta,.90}$	$\eta = \tau_{\eta,.75}$	$\eta = \tau_{\eta,.90}$
(e) $\rho_c = 0.01$	1.4	2.2	4.2	7.9	7.6	14.4
(f) $\rho_c = 0.02$	1.3	1.9	3.5	6.5	6.2	11.8
(g) $\rho_c = 0.05$	1.1	1.5	2.6	4.6	4.5	8.4
(h) $\rho_c = 0.20$	1	1.2	1.6	2.6	2.5	4.6

Notes: For each combination of intracluster correlation (ρ_c), intercluster correlation (ρ_a), number of clusters in each small area (C), and area fixed effect (η), the figure represents the ratio between the (correct) standard error of the prediction error of mean expenditure and the (incorrect) standard error calculated assuming $\rho_a = \eta = 0$. All calculations assume that each cluster contains 100 households. Given a probability distribution function $f(\eta)$ for the area fixed-effect η, the value $\tau_{\eta,p}$ is the pth percentile of the distribution, so that $P(\eta \leq \tau_{\eta,p}) = p$. The coefficients β in the conditional expectation of y are assumed to be known.

including a large number C of clusters labeled c, each of which includes a population of m households labeled h. For simplicity, and for this section only, we assume that both C and m are constant and that the welfare measure of interest is mean expenditure in area a, which we denote by μ_y^a. We also assume an equicorrelated structure for the errors, and treat the area fixed effect as random, even if the specific value of the fixed effect η is treated as a constant for a given small area. Specifically:

$$\text{Var}(u_{ach}) = \sigma_u^2$$

$$\text{Cov}(u_{ach}, u_{a'c'h'})$$

$$= \begin{cases} 0 \text{ if } a \neq a' \\ \quad \text{(no correlation between areas)} \\ \sigma_a^2 = \rho_a \sigma_u^2 \text{ if } a = a', c \neq c' \\ \quad \text{(same area, different clusters)} \\ \sigma_c^2 = \rho_c \sigma_u^2 \text{ if } a = a', c = c', h \neq h' \\ \quad \text{(same cluster, different household),} \end{cases}$$

where ρ_a and ρ_c are, respectively, the intra-area (intercluster) and the intracluster correlation coefficients. In the specific case where the error term has a random effects structure as in (6), the total variance of the error is $\sigma_u^2 = \sigma_\eta^2 + \sigma_e^2 + \sigma_\varepsilon^2$, while $\rho_a = \sigma_\eta^2/\sigma_u^2$ and $\rho_c = (\sigma_\eta^2 + \sigma_e^2)/\sigma_u^2$. We are particularly interested in the consequences of assuming area homogeneity, as in the standard poverty mapping exercise, which here means assuming that $\sigma_\eta^2 = 0$ (implying $\rho_a = 0$) when it is not in fact true. We also assume that β is known, so that our argument will abstract from the existence of estimation error in these parameters. Note that

this estimation error will contribute to the MSE of estimation of μ_y^a, whether or not homogeneity holds.

The estimator for the mean expenditure in a given small area A will be

$$\hat{\mu}_y^a = \frac{1}{Cm} \sum_{c=1}^{C} \sum_{h=1}^{m} X'_{ach}\beta,$$

so by using the structure of the error term, the MSE can be written as

$$MSE = E[(\hat{\mu}_y^a - \mu_y^a)^2 | a = A] = \eta_A^2 + \frac{\sigma_e^2}{C} + \frac{\sigma_\varepsilon^2}{Cm}$$

$$= \eta_A^2 + \frac{\sigma_u^2}{Cm}\left[(\rho_c - \rho_a)m + (1 - \rho_c)\right] \qquad (7)$$

$$= \eta_A^2 + \frac{\sigma_u^2}{Cm}\left[1 + \rho_c(m - 1)\right] - \frac{\sigma_\eta^2}{C}.$$

The second term coincides with the variance of the estimator when the DGP in (6) does not include an area fixed effect, so that $\eta_A = \sigma_\eta = 0$. Both this and the third term converge to 0 when the number of clusters in the small area becomes large, but the first term does not, and can lead to severe underestimation of the MSE in areas characterized by a nonzero value of η_A.

Table 1 shows the underestimation of the root MSE for a given small area that would result from incorrectly assuming that area fixed effects are 0. We tabulate results for

different parameter combinations, keeping cluster size fixed at $m = 100$.[5] Each figure is the ratio between the (true) root MSE calculated as in (7) and the incorrect root MSE calculated assuming $\rho_a = \sigma_\eta = 0$, which is given by the second term in (7). For each combination of ρ_c, ρ_a, and C, we calculate ratios for two different values of the area fixed effect η, which are taken to be the 75th and 90th percentile of the distribution of η. We assume that the distribution of u is normal with mean zero and unit variance (it is straightforward to check that the unit variance is simply a choice of units); given ρ_c, ρ_a, and C, σ_η^2 and σ_ε^2 are set, as is the distribution of η.

The results show that disregarding the bias component can lead to severe underestimation of the MSE even when the small-area fixed effect is small and even when the intracluster correlation is below 0.05 or lower. For example, take the case where each area encompasses 150 clusters, the intracluster correlation is 0.01, and $\rho_a = 0.005$. For a small area whose fixed effect is equal to the 75th percentile of the distribution of η (row e, column 3) the ratio between correct and naive MSE is 4.2, which also means that the ratio will be even larger for the 50% of the small areas whose absolute value of η is larger than the 75th percentile. Given the same DGP, the correct MSE will be at least 7.9 times larger than the naive one for 20% of small areas (row e, column 4). The relative underestimation of the MSE generally worsens if the number of clusters within a small area increases and becomes smaller if the intercluster correlation becomes small relative to the intracluster correlation. Overall, the ratios in the table range from 1 to 19.9, both resulting from unlikely combinations that require a very high intracluster correlation equal to .20.

The MSE in (7) is calculated conditional on a specific area effect η_A. We are also interested in the unconditional MSE for μ_y integrated over the distribution of η. In this case, the underestimation of the MSE from ignoring the heterogeneity is closely analogous to the underestimation of standard errors that comes from ignoring the complex survey design of household survey data. Appendix A shows that the unconditional MSE, which here coincides with the sampling variance of $\hat{\mu}_y$, can be written as:

$$
\text{Var}(\hat{\mu}_y) = \left(\frac{\sigma_u}{\sum_{c=1}^C m_c} \right)^2
$$

$$
\times \left\{ \underbrace{\sum_{c=1}^C m_c}_{} + \underbrace{\sum_{c=1}^C m_c(m_c - 1)\rho_c}_{\text{from intracluster corr.}} + \underbrace{\sum_{c=1}^C \sum_{c'=1,c'\neq c}^C m_c m_{c'}\rho_a}_{\text{from intercluster corr.}} \right\}. \tag{8}
$$

The first term (in larger braces) is the variance calculated assuming that observations are independent and identically distributed (i.i.d.). The second and third terms come, respec-

tively, from the intracluster and intercluster correlation implied by model (6), because of the common geographical and socioeconomic characteristics within the area that come from the failure of area homogeneity. This last term can be large. In the simple case where each cluster contains the same number of households, so that $m_c = m \forall c$, equation (8) simplifies to

$$
\text{Var}(\mu_y - \hat{\mu}_y)
$$

$$
= \frac{\sigma_u^2}{Cm} [1 + (m - 1)\rho_c + m(C - 1)\rho_a]
$$

$$
= \text{Var}_{SRS} + \frac{\sigma_u^2}{Cm} [(m - 1)\rho_c + m(C - 1)\rho_a] \tag{9}
$$

$$
= \text{Var}_C + \frac{\sigma_u^2}{Cm} [m(C - 1)\rho_a],
$$

where Var_{SRS} is the variance estimated under the assumption of i.i.d. observations and Var_C is the variance estimated under the assumption that observations are correlated within clusters but independent across clusters. Although Var_C goes to 0 as the number of clusters goes to infinity, the second term in the last line converges to $\rho_a \sigma_u^2$, which is not 0 unless the intra-area (intercluster) correlation ρ_a is 0. In consequence, even if ρ_a is small, the ratio of the correct MSE to the Var_C, which is the MSE ignoring the intracluster correlation, goes to infinity with C. Even with $C = 150$, $m = 100$, and an intercluster coefficient of only 0.01, the ratio of the correct to incorrect root MSE is 2.9 when the intracluster coefficient is 0.20, is 5.1 when it is 0.05, and is 7.1 when it is 0.02, so that the variance is underestimated fifty-fold.

These unconditional results, as well as the conditional results in table 1, exaggerate the practical effects of ignoring intercluster correlation because they exclude the contribution to the MSE of estimating the β parameters, a contribution that is common to both the correct and the incorrect MSE and whose inclusion would bring their ratio toward unity. In the other direction, we have so far maintained the assumption that there is no inter-area variation in β. As we shall see in section VB, violation of this condition will also have an impact on the MSE.

IV. Estimation

In this section we describe a simple parametric estimator for the estimation of poverty maps together with a brief description of the more complex methodology that Elbers et al. proposed that is routinely used in poverty mapping.

A. A Simple Projection-Based Estimator

We assume that both MP and AH hold. Given AH, we can see from the modified moment condition (5) that the sampling process identifies the parameter of interest, and we

propose an estimator based on the simple idea of replacing the modified moment condition (5) by its sample analog. The estimate $\hat{W}_0$ is then obtained as the solution to the following equation:

$$\frac{1}{N_A} \sum_{h \in H(A)} \hat{E}[s_h g(y_h; \hat{W}_0)|X_h] = 0, \tag{10}$$

where N_A is total population in the small area according to the census, and the expectation can be approximated by a projection of $s_h g(y_h)$ on a series of functions of X_h.[6]

To fix ideas, suppose that the welfare measure of interest W_A is the head count poverty ratio in a small area A, calculated for a given fixed poverty line z. If we disregard for simplicity the difference between household and individual-level data, the parameter of interest then becomes

$$W_A = \frac{1}{N_A} \sum_{h \in H(A)} 1(y_h \leq z). \tag{11}$$

Under assumptions MP and AH, the head count can be estimated in two steps. First, the parameters γ that describe the conditional probability $P(y_h \leq z|X_h; \gamma)$ are estimated with a parametric binary dependent variable model such as logit or probit using survey data from region R. Then the poverty count is estimated as

$$\hat{W}_A = \frac{1}{N_A} \sum_{h \in H(A)} P(y_h \leq z|X_h; \hat{\gamma}), \tag{12}$$

that is, as the mean of the imputed probabilities over all census units from area A. In the rest of the paper, we will refer to (12) as the projection estimator.

Because the census population is kept fixed and is therefore nonrandom, the only source of sampling error in (12) is the estimation of the parameters γ so the standard error can be calculated using the delta method (see, e.g., Wooldridge, 2002a). If a logit model is adopted in the first stage, the delta method leads to

$$\widehat{\text{Var}}(\hat{W}_A) = \hat{G}\widehat{\text{Var}}(\hat{\gamma})\hat{G}', \tag{13}$$

where

$$\hat{G} \equiv \frac{1}{N_A} \sum_{h \in H(A)} \frac{e^{\tilde{X}_h'\hat{\gamma}}}{(1 + e^{\tilde{X}_h'\hat{\gamma}})^2} \tilde{X}_h',$$

where $\tilde{X}_h$ denotes the covariates used to estimate the projection and $\widehat{\text{Var}}(\hat{\gamma})$ is the estimated covariance matrix of the first-stage coefficients, calculated taking into account the clustered survey design.[7]

Note, however, that the MSE of the estimator should take into account not only the variance of $\hat{W}_A$ but also the difference between W_A as defined in (11) and the census mean of $P(y_h \leq z|X_h; \gamma)$. This difference, which we refer to as a bias, would be present in the estimation of W_A even if γ were known. In Appendix B, we show that the contribution of this bias to the MSE can be approximated by

$$\hat{b}^2(\hat{W}_A) = \frac{\hat{E}[(p_h - 1(y_h < z))^2]}{N_A} + \frac{N_A - 1}{N_A} \tag{14}$$
$$\times \hat{E}[(p_h - 1(y_h < z))(p_{h'} - 1(y_{h'} < z))],$$

where $p_h = P(y_h \leq z|X_h; \gamma)$ and the second expectation on the right-hand side is taken with respect to different households within the same cluster. Both expectations can be estimated using their respective sample analogs. To summarize, a confidence interval for $\hat{W}_A$ with nominal coverage $(1 - \tau)$ will be constructed as

$$\hat{W}_A \pm \Phi^{-1}(1 - \tau/2) \times [\widehat{\text{Var}}(\hat{W}_A) + \hat{b}^2(\hat{W}_A)], \tag{15}$$

where $\Phi^{-1}(\cdot)$ is the inverse of the cumulative distribution function of a standard normal. It should be noted that while the first term in (14) goes to 0 when the size of the small area increases, the second term does not. Therefore, the bias component may be relatively large when the residuals of the first-stage model exhibit a large intracluster correlation, because such correlation will lead to a large second term. Appendix B contains a Monte Carlo simulation that confirms this finding and shows that, as is to be expected, the bias adjustment is particularly important when the area is small.

Adapting this approach to the estimation of parameters other than poverty head counts is relatively straightforward as long as W_A can be written as a function of parameters identified by a moment condition. For instance, if W_A is the poverty gap, $g(y_h; \hat{W}_A) = 1(y_h \leq z)(1 - y_h/z) - W_A$, so that in the first stage, one can estimate $E[1(y_h \leq z)(1 - y_h/z)|X_h]$ by projecting the poverty gaps on a list of functions of X_h. The parameter of interest can then be calculated as the mean predicted value for all census units and the standard errors calculated using the delta method as in (13), with $G = X_h$. The estimation of the squared bias would also proceed in an analogous way.

B. The ELL Estimator

The poverty maps constructed by the World Bank or with their assistance make use of an alternative estimation

---

[6] Note that this approach also lends itself well to nonparametric estimation, as in Chen et al. (2005) and Chen et al. (2008). However, the accurate implementation of nonparametric estimation requires the choice of smoothing parameters, and it may be cumbersome when the number of predictors is large, so we do not pursue this direction further. Even so, note that parametric rates of convergence of the parameter of interest can still be achieved because W_0 is calculated as the integral of a conditional expectation and is not a conditional expectation itself.

[7] The variables X and $\tilde{X}$ do not necessarily coincide, because the latter may include, for instance, powers or interactions.

method proposed by Elbers et al. (2003) (ELL for brevity). Like our estimator, ELL requires the validity of both MP and the area homogeneity assumption AH. Unlike the estimator described in the previous section, ELL is a simulation-based estimator that requires explicit parametric assumptions about the distribution of regression residuals. In the context of poverty mapping, we will need functional forms for the conditional mean of y, as well as for the distribution around that mean. Different variants of ELL have been described in the literature, but all of them share the same central features. As a basis for the calculations below, we provide a brief description (for more, see Elbers et al., 2002; Elbers, Lanjouw, & Lanjouw; 2003; Demombynes et al., 2007). Expenditure for household h in EA c is modeled as

$$\ln(y_{ch}) = \beta'X_{ch} + u_{ch} = \beta'X_{ch} + \eta_c + \varepsilon_{ch},$$

where $\text{Cov}(\eta_c, \varepsilon_{ch}) = \text{Cov}(X_c, \eta_c) = \text{Cov}(X_c, \varepsilon_{ch}) = \text{Cov}(\varepsilon_{ch}, \varepsilon_{c'h'}) = 0 \; \forall \; c, c', h, h' \neq h$. The idiosyncratic errors are allowed to be heteroskedastic, while the cluster fixed effects are assumed to be i.i.d. and homoskedastic; higher conditional moments are not considered. Consistent estimation of β is clearly not sufficient for the estimation of poverty or inequality measures, which are functions of the distribution of y, and not functions of the distribution of the conditional expectation $\beta'X_{ch}$. For this reason, once β has been estimated using ordinary least squares or feasible generalized least squares, ELL use a simulation procedure to recreate the conditional distribution of y by adding to each estimated fitted value $\hat{\beta}'X_{ch}$ simulated values of the cluster-specific (η_c) and household-specific (ε_{ch}) errors. Because the errors u_{ch} are not i.i.d., the simulated draws must take into account the clustering and heteroskedasticity. Several alternative algorithms have been proposed for this; all start from the separate estimation of η_c and ε_{ch}. Once $\hat{\beta}$ has been obtained, the cluster fixed effects are estimated as the mean value of the residuals $\hat{u}_{ch}$ over all the observations from the same cluster c. Estimates e_{ch} of the idiosyncratic errors are then calculated as $\hat{u}_{ch} - \hat{\eta}_c$. The variance of the idiosyncratic error ε_{ch} is then estimated imposing the following parametric form for heteroskedasticity:

$$\sigma^2_{\varepsilon,ch}(X) = \frac{Ae^{z'_{ch}\alpha} + B}{1 + e^{z'_{ch}\alpha}}, \tag{16}$$

where z_{ch} is a function of the correlates X, and A and B are parameters to be estimated. When the estimates from (16) are used, standardized residuals are then calculated as

$$e^*_{ch} = \frac{e_{ch}}{\hat{\sigma}_{\varepsilon,ch}} - \frac{1}{H}\sum_{ij}\frac{e_{ij}}{\hat{\sigma}_{\varepsilon,ij}}.$$

The point estimates and corresponding variances of β and the heteroskedasticity parameters, together with the empirical distribution of the cluster-specific and idiosyncratic errors, are the inputs that can now be used to estimate W_0 and its standard error.

The structure of each simulation step r is as follows. First, a set of parameters is drawn from the sampling distribution of β and the parameters in (16). Second, each cluster in the census is assigned a cluster-specific error $\hat{\eta}_c^r$ drawn from the empirical distribution of all $\hat{\eta}$. Third, each observation in the census is assigned a normalized idiosyncratic error e_{ch}^{*r}, which is obtained from either a parametric distribution or the empirical distribution of the errors. Fourth, heteroskedastic errors e_{ch}^r are calculated by using the parametric model in (16) evaluated at the simulated parameter values. Finally, simulated values for $\ln y$ are generated as $\ln y^r = \hat{\beta}^{r'}X_{ch} + \hat{\eta}_c^r + e_{ch}^r$, and a value W^r is then simply calculated based on the simulated expenditure data. The mean and the variance over a large number of simulations are then used as an estimate of $\hat{W}$ and $\widehat{\text{Var}(\hat{W})}$ (note the similarity with multiple imputation; Rubin, 1987). The bias adjustment described in section IVA is not necessary in ELL, because the simulation procedure accounts automatically for the presence of the bias.[8]

This approach disregards the possible correlation among observations that belong to different clusters and will therefore overstate the precision of the estimates if such correlation exists. Such would be the case, for instance, if the true model includes area fixed effects such as in (6). Elbers et al. (2002) argue that in such cases, one can modify ELL to obtain an upper bound of the true variance: in each replication, instead of assigning the same location effect estimated at the cluster level to all units within a cluster, one can assign the location effect to all units within the same area. Alternatively, one can also experiment attaching the estimated cluster effects to geographical levels intermediate between the cluster and the area. This would lead to larger and larger (and possibly more and more conservative) standard errors the closer to the area is the chosen level of aggregation. Elbers et al. (2002) argue that when the intracluster correlation is small, such conservative estimates of the standard errors will be only marginally different from those that assume no intercluster correlation, but they explore only the consequences of correlation with locations smaller than the area of interest. The arguments laid out in section IIIA, as well as results from Monte Carlo experiments in the next section, suggest that such an upper bound can be very much larger than the standard errors calculated under the assumption of no intercluster correlation even in cases where intracluster correlation is very small. The very large standard errors that may arise from the use of such conservative estimates may be one reason that their use appears to have been ignored in poverty mapping (see, e.g., Mistiaen et al., 2002; Alderman et al., 2003; Elbers et al., 2004, 2007; Demombynes & Õzler, 2005). Even if the actual consequences of intra-area correlation, as we have

---

[8] ELL define the bias contribution to the MSE of their estimator as the "idiosyncratic" component of the standard error.

TABLE 2.—MONTE CARLO SIMULATIONS: NO INTERCLUSTER CORRELATION

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
			Projection Estimator			ELL	
	True Value	Bias	RMSE	Coverage	Bias	RMSE	Coverage
$P_0(24)$	0.0979	0.0007	0.0092	0.976	0.0015	0.0079	0.988
$P_0(25)$	0.3323	−0.0026	0.0148	0.984	−0.0031	0.0129	0.972
$P_0(26)$	0.6732	−0.0053	0.0150	0.964	−0.0056	0.0128	0.968
$P_1(24)$	0.023	−0.0000	0.0003	0.940	0.0000	0.0003	0.972
$P_1(25)$	0.0103	0.0000	0.0006	0.972	0.0001	0.0006	0.988
$P_1(26)$	0.0292	−0.0000	0.0009	0.980	0.0000	0.0010	0.988

Note: 250 Monte Carlo replications. The synthetic census population is composed of 150 enumeration areas of 100 households each. The sample drawn in each replication includes 1,000 households selected from 100 equally sized clusters. The bias is calculated as the mean deviation of the estimates from the true value over the 250 simulations. The RMSEs are the squared roots of the same deviations squared. Coverage rates are calculated for 95% confidence intervals.

shown, depend on several factors, it would be useful if conservative estimates were routinely produced and evaluated. We also note that these considerations address only the existence of location fixed effects, while ignoring any heterogeneity in the slopes of the estimated conditional model.

ELL's simulation estimator has the advantage of allowing the estimation of any poverty or inequality measure within the simulation procedure used for estimation. After a replication has generated a complete census of expenditures y, any welfare measure can be easily calculated using the generated y as if they were data. This works even for measures such as the Gini coefficient that are not identified by a simple moment condition (see Section I). But this versatility comes at the price of parametric assumptions about the conditional mean of y, its conditional variance, and the absence of conditional skewness or kurtosis, for example. The estimator we have described in (12), while still parametric, models only the conditional probability of being in poverty. In this way, the estimator is faster and simpler, and it does not require assumptions on the complete conditional distribution or first-stage residual errors. Both estimators, however, require the absence of area heterogeneity, and this common ground is likely more important than their differences.

V. Monte Carlo Experiments

We first consider a best-case scenario where the DGP is characterized by a simplified version of model (6), where there is no small-area fixed effect, the cluster fixed effects are i.i.d. and homoskedastic, and MP and AH hold. Specifically, for each cluster within a region, the DGP is described as follows:

$$y_{ch} = \beta_0 + \beta_1 x_{ch} + u_{ch} = 20 + x_{ch} + e_c + \varepsilon_{ch}$$

$$x_{ch} = 5 + z_{c,1} w_{ch} + z_{c,2}, \quad w \sim N(0, 1),$$

$$z_{c,1}, z_{c,2} \sim U(0, 1), \quad z_{c,1} \perp z_{c,2}, \tag{17}$$

$$e_c \sim N(0, .01), \quad \varepsilon_{ch} \sim N(0, \sigma^2(x))$$

$$\sigma^2(x) = \frac{e^{\alpha_1 x + \alpha_2 x^2}}{1 + e^{\alpha_1 x + \alpha_2 x^2}},$$

with $\alpha_1 = .5$, $\alpha_2 = -.01$. The idiosyncratic errors ε_{ch} are then assumed to be heteroskedastic, and their variance is determined by a simplified version of model (16) so that the functional form of the heteroskedasticity is consistent with the assumptions in ELL. This model implies that the proxy variable x explains approximately 30% of the variance of y. The intracluster correlation coefficient, calculated as $\sigma_e^2/(\sigma_e^2 + \sigma_\varepsilon^2)$, is small and approximately equal to .027.

One conceptual complication in performing a Monte Carlo (MC) experiment in this context is that the population of interest (synthetic households in a small area) is finite and relatively small (e.g., 15,000 to 20,000 households), and the quantity to be estimated (e.g., a poverty ratio) is itself a function of this finite population rather than being a fixed parameter as in a typical MC simulation. In our case, the DGP described would generate a unique value of a welfare measure only in a population composed of an infinite number of EAs and households, but in assessing the performance of different estimators, we think it is important to work with a population of size analogous to the ones met in real empirical applications with census data. Hence, we use the DGP to generate a population of $N_A = 15,000$ households divided into 150 EAs of 100 households each. This population represents the small area A for which a welfare indicator has to be calculated. We assume that the researcher is interested in estimating head count poverty ratios, $P_0(z)$, and poverty gaps, $P_1(z)$, evaluated at three different poverty lines $z = 24, 25, 26$. The true values of the six poverty measures in the artificial population are reported in column 1 of table 2. In each MC replication, we use the DGP to generate an artificial sample of 10 households from each of 100 randomly generated clusters. For simplicity, we ignore the fact that a few observations in the auxiliary sample may belong to the same small area of interest. Because the usefulness of the estimation approaches considered in this paper hinges on the fact that the number of such observations is typically very small, the correlation should be of little or no consequence in the calculation of the standard errors.

For each estimator we calculate bias, RMSE, and confidence interval coverage rates ("coverage" in the table) for 95% nominal coverage rates intervals. Bias is calculated as

$R^{-1} \sum_{r=1}^{R} (\hat{W}_A^r - W_A)$, where W_A is the true value of the welfare measure, and $\hat{W}_A^r$ is the estimate obtained in the rth MC replication. The RMSE is estimated as the square root of $R^{-1} \sum_{r=1}^{R} (\hat{W}_A^r - W_A)^2$, and coverage rates are calculated as the fraction of the replications for which the true value lies within a 95% nominal confidence interval.

We consider the performance of the projection and simulation-based ELL estimators described in section IV. All MC replications use the same artificial census population, which is therefore treated as nonrandom. For a given auxiliary sample generated in the rth replication, we calculate the projection estimator as described in the previous section, using as predictors x and its square, $\sin(x)$, $\cos(x)$, $\sin(2x)$, and $\cos(2x)$. In empirical applications, the degree of flexibility in the choice of the functional form will be limited by the number of predictors and the size of the survey sample. When adopting the ELL estimator, we estimate the heteroskedasticity parameters (α_1, α_2) using nonlinear least squares, using the correct model (17) described in the DGP. At each step of the ELL procedure, two sets of parameters, (β_1, β_2) and (α_1, α_2), are drawn from their respective estimated asymptotic distributions. Each EA in the artificial census is then assigned a cluster-specific fixed effect drawn at random (with replacement) from the set of all fixed effects estimated as described in section IVB. The household-specific standardized fixed effects are similarly assigned to each unit after being randomly selected with replacement from the empirical distribution of all e_{ch}^*, and then transformed into heteroskedastic errors using the random draw of the heteroskedasticity parameters.

Table 2 reports the results of 250 MC replications. For all welfare measures, both estimators are essentially unbiased, the RMSE is small relative to the true value being estimated, and coverage rates are very close to the nominal 95%. Overall, when the parametric assumptions ELL use are correct, both estimators perform well, although the projection estimator is substantially simpler than ELL. Both estimators, however, rely on the absence of heterogeneity across areas within the same region. In the next section we explore the consequences of the failure of this assumption, which we deem likely to arise with real data.

A. Consequences of Heterogeneity on Coverage Rates

We first consider the case where the true DGP for expenditure includes not only an EA fixed effect but also a small-area fixed effect, as in (6). We assume that there is no heterogeneity in the slopes β, an assumption that we will relax in section VB. For simplicity, we also assume homoskedastic errors. The DGP for expenditure of household h is cluster c in small area a is now assumed to be described by the following:

$$y_{ach} = 10 + 2x_{ach} + \eta_a + e_{ch} + \varepsilon_{ach}, \qquad (18)$$

$$x \sim N(5, 1) \quad \eta_a \sim N(0, \sigma_\eta^2)$$
$$e \sim N(0, \sigma_e^2) \quad \varepsilon \sim N(0, \sigma_\varepsilon^2). \qquad (19)$$

Note that in this case, welfare estimates will depend on the area fixed effect η_a. For instance, letting z denote a fixed poverty line, the head count poverty ratio in a given small area a becomes

$$P(y_{Ach} \leq z | A = a) = P(2x + e_{ch} + \varepsilon_{ach} \leq z - 10 - \eta_a)$$
$$= \Phi\left(\frac{z - 10 - \eta_a}{\sqrt{\sigma_e^2 + \sigma_\varepsilon^2 + 4}} \right),$$

where the last expression follows from the normality of the errors and of the covariate x. As in the previous MC, we consider small areas of 15,000 households split among 150 equally sized EAs. However, to introduce heterogeneity in the population, we assume that the region from which the auxiliary data set is drawn is composed of 25 small areas characterized by the same distribution of x but different values of the area fixed effect. To avoid the (albeit unlikely) possibility of an unusual draw of area fixed effects, which are assumed to be generated from a normal distribution with variance σ_η^2, we set the 25 area fixed effects equal to τ_{η,p_i}, $i = 1, \ldots, 25$, $p_i = 0.01 + (.98/24) \times (i - 1)$, where τ_{η,p_i} is the p_i-quantile of the assumed distribution of η, so that $P(\eta \leq \tau_{\eta,p_i}) = p_i$. Hence, for instance, when $i = 1$, the area has η_a equal to the first percentile of the assumed distribution; when $i = 13$, the fixed effect is equal to 0 (the median and mean of the distribution); and when $i = 25$, it is equal to the 99th percentile. When the area fixed effect becomes larger, we should expect the performance of both estimators to worsen, with coverage rates decreasing towards 0.

Monte Carlo results for three alternative models based on 200 replications are displayed in table 3. Each ELL estimation is obtained with 200 simulations. In each model, we keep $\sigma_\varepsilon = 2$ while we experiment with different values of σ_η and σ_e. In each replication, an artificial sample of 1,000 households is generated from the DGP in (18) and (19). We draw four EAs from each of the 25 small areas, and then we draw 10 households from each EA. For each DGP, the object of interest is the head count ratio calculated for a poverty line corresponding to the 25th percentile of the overall distribution of y in the whole region. In both models, the predictor x has good explanatory power, with an R^2 approximately equal to .50.

Because the DGP assumes that the errors are homoskedastic, the ELL estimator can take a simpler form than described in section IVB. The first step is unchanged and consists of the estimation of model (18) using ordinary least squares, followed by calculation of the empirical distribution of cluster-specific and idiosyncratic residuals. At each simulation, an intercept and a slope are drawn from their respective estimated sampling distributions. Then each EA

TABLE 3.—MONTE CARLO SIMULATIONS: CONSEQUENCES OF INTERCLUSTER CORRELATION

	P_0	η_a	(1) Bias	(2) SE	(3) RMSE	(4) Coverage	(5) Bias	(6) SE	(7) RMSE	(8) Coverage	(9) Bias	(10) SE	(11) RMSE	(12) Coverage
			Projection Estimator				ELL				ELL, Conservative SE			
Model 1: $\sigma_\eta = .5$, $\sigma_e = .5$, $\sigma_\varepsilon = 2$. $R^2 = .471$, $\rho_c = .111$, $\rho_a = .056$														
(1A)	0.246	$\tau_{\eta,.50}=0$	0.00444	0.01250	0.01324	0.905	0.00588	0.01038	0.01190	0.980	0.00738	0.10615	0.01503	1
(1B)	0.212	$\tau_{\eta,.75}= .329$	0.03908	0.01250	0.04103	0.090	0.04059	0.01038	0.04189	0.085	0.04208	0.10615	0.04407	1
(1C)	0.166	$\tau_{\eta,.95}= .818$	0.08486	0.01250	0.08577	0.000	0.08646	0.01038	0.08708	0.000	0.08796	0.10615	0.08893	1
Model 2: $\sigma_\eta = .25$, $\sigma_e = .10$, $\sigma_\varepsilon = 2$. $R^2 = .496$, $\rho_c = 0.178$, $\rho_a = .0153$														
(2A)	0.249	$\tau_{\eta,.50}=0$	0.00235	0.01120	0.01141	0.960	0.00391	0.00856	0.00939	0.975	0.00487	0.07810	0.01157	1
(2B)	0.231	$\tau_{\eta,.75}= .165$	0.02041	0.01120	0.02327	0.560	0.02198	0.00856	0.02358	0.530	0.02294	0.07810	0.02522	1
(2C)	0.205	$\tau_{\eta,.95}= .409$	0.04582	0.01120	0.04716	0.005	0.04739	0.00856	0.04816	0.000	0.04835	0.07810	0.04948	1
Model 3: $\sigma_\eta = .10$, $\sigma_e = .25$, $\sigma_\varepsilon = 2$. $R^2 = .496$, $\rho_c = .0178$, $\rho_a = .002$														
(3A)	0.249	$\tau_{\eta,.50}=0$	0.00128	0.01131	0.01136	0.935	0.00288	0.00907	0.00949	0.975	0.00409	0.07779	0.01162	1
(3B)	0.242	$\tau_{\eta,.75}= .066$	0.00858	0.01131	0.01417	0.885	0.01018	0.00907	0.01362	0.890	0.01139	0.07779	0.01575	1
(3C)	0.231	$\tau_{\eta,.95}= .164$	0.01920	0.01131	0.02227	0.655	0.02080	0.00907	0.02268	0.530	0.02201	0.07779	0.02456	1

Note: All results are based on 200 Monte Carlo replications. The simulation-based results in columns 5–12 use 150 simulations within each Monte Carlo replication. A synthetic small area includes 150 EAs with 100 households each. Samples have size 1,000 and are generated drawing 4 EAs from each small area and 10 households from each EA. Coverage rates are calculated for confidence intervals with nominal coverage equal to .95. RMSE denotes root mean squared error. $\tau_{\eta,p}$ is the pth quantile of the distribution of η_a. The standard errors (*) in column 10 are calculated as the mean estimated standard errors (calculated over all Monte Carlo simulations) when the same cluster fixed effect is added to all observations within a small area. The poverty line corresponds to the 25th percentile of the overall distribution of y in the whole region. See section VA for simulation details.

and each household in the artificial population is matched to a cluster-specific fixed effect drawn at random (with replacement) from the corresponding empirical distribution, while no adjustment for heteroskedasticity is necessary in this case. We obtain the results for the projection estimator using the same two-step methodology used for poverty head counts in the previous MC.

The top half of table 3 shows a first set of results, where the DGP implies moderately large intracluster correlation (.11) and intercluster correlation (.06). Both the projection estimator (columns 1–4) and ELL (columns 5–8) perform well in predicting the poverty counts when the area fixed effect is 0 in terms of both bias and coverage (row 1A). The performance of both estimators worsens considerably when a small-area fixed effect is present. If the small area includes a fixed effect equal to .329 (the 75th percentile of the distribution of η and less than 2% of the mean value of the "expenditure" variable y), the coverage rate for both estimators remains below 10% and is actually very close to 0 in several cases (row 1B, columns 4 and 8). The results in row 1C show that when the area fixed effect is .818 (the 95th percentile of the distribution of η), the coverage rate decreases to 0 for both estimators. Consistently with the results in section III, the decline in coverage is caused by the increase in bias associated with the presence of the area fixed effect. While in row 1A virtually all MSE derives from the standard error of the estimator (because the bias is close to 0), in rows 1B and 1C, the standard error becomes only a fraction of the MSE, and so the estimated confidence intervals provide misleading information about the true poverty head counts.

In columns 9 to 12 we show results obtained again with ELL but calculating the standard errors using the conservative approach described in section IVB. In this case, in each

of the 150 simulations required to complete one of the 200 Monte Carlo replications, we assign the same cluster fixed effect to all households within the same area. This modification should lead to a very conservative confidence interval, because it assumes that the correlation between two units from two different EAs within the same cluster is the same as the correlation between two units from the same EA.[9] In fact, this modified methodology leads to standard errors that are approximately ten times as large as those estimated in column 6. The increase in the standard errors now leads to confidence intervals that always include the true value, so that coverage rates are equal to 1 in all cases, even when the area fixed effect is relatively large (row 1C). However, the confidence intervals are now so wide as to become barely informative. For instance, standard ELL produces a confidence interval of width 0.041 (0.0104 × 1.96 × 2), while conservative ELL produces intervals of width .416 (0.10615 × 1.96 × 2).

In rows 2A to 3C of table 3, we show that coverage rates may be far from the nominal 95% even in cases where intracluster correlation is very small. The DGP in model 2 implies a small intracluster correlation (.0178), but also implies that most of it is due to the presence of the area fixed effect, so that the intercluster correlation is .0153. As a result, coverage rates decline rapidly for both ELL and the projection estimator, and when the area fixed effect is moderately large (row 2C), coverage approaches or equals 0. This result is consistent with the figures in table 1, where we have shown that when the ratio between inter- and intracluster correlation is large, standard errors that do not

[9] This also explains why the mean conservative standard errors reported in column 10 of table 3 are larger than the RMSE. The latter is the sum of the true bias and variance as calculated across MC simulations.

take intercluster correlation into account will seriously underestimate the true MSE, leading to misleading inference. This is also confirmed in the results reported in the last three rows of the table, where the intracluster correlation is the same as in model 2, but the intercluster correlation is close to 0 (.002). In this case, for areas where the fixed effect is as large as the 75th percentile of the distribution of η, coverage rates remain close to the nominal 95% level. When the area fixed effect is as large as the 95th percentile, coverage decreases further but remains above 50%, and the bias is very small.

Overall, disregarding the presence of area fixed effects may lead to severely misleading inference even when the intracluster correlation accounts for less than 2% of the total variance of the error, unless the area fixed effects are very small relative to the EA fixed effects. The overstatement of precision can be repaired by using ELL's conservative standard errors, but this generates confidence intervals that are so wide as to render the estimates barely informative. Note also that we have assumed that unobserved heterogeneity is present only in the intercept, while the slopes are the same for all observations that belong to the same region. In empirical applications, it is likely that inference will be complicated by heterogeneity in the slopes that link the predictors to expenditure, for example, by spatial variation in the rates of return to physical and human capital. In the next section, we explore how bias, MSE, and coverage rates are affected when heterogeneity in the expectation of y given X takes the form of heterogeneity in the slopes.

B. A Model with Heterogeneity in Slopes

We now assume that the following random coefficient model holds:

$$y_{ach} = 10 + \beta_a x_{ach} + e_c + \varepsilon_{ach},$$ (20)

$$x \sim N(5, 1), \quad \beta_a \sim N(3, \sigma_\beta^2), \quad \sigma_e = .1, \quad \sigma_\varepsilon = 3.$$

This model can be rewritten as a model with homogeneous slopes and heterogeneity in the heteroskedasticity. The DGP is indeed equivalent to one with $y_{ach} = 10 + \beta x_{ach} + e_c + \theta_a x_{ach} + \varepsilon_{ach}$, where $\theta_a \sim N(0, \sigma_\beta^2)$. Note also that by construction, this form of heterogeneity generates intercluster, and therefore intracluster, correlation.

As in the previous section, we assume that survey data are available from a region composed of 25 areas and that the population of each area is divided into 150 EAs of 100 households each. The area-specific slopes β_a are kept constant throughout all MC replications. As before, we select the slopes using a semirandom approach by setting them equal to $\sigma_\beta \Phi^{-1}(q_i)$, $i = 1, 2, \ldots, 25$, where $\Phi(\cdot)$ is the cumulative distribution function of a standardized normal and the 25 q_is are the bounds of 24 equally sized subsets of the interval 0.01–0.99. This procedure ensures that even if the slopes are kept constant and are selected according to a deterministic procedure, their value can be thought of as a representative draw from the distribution of β_a described in (20). Note also that by construction, one of the small areas is characterized by a slope equal to the median (mean) of the distribution of β_a.

In each MC replication, a sample of 500 units is generated from the DGP in (20), drawing one cluster of 20 observations from each of the 25 areas. We consider the estimation of $P_0(20)$, $P_0(22)$, and $E(y)$ in two cases with different degrees of heterogeneity in the slope ($\sigma_\beta = 0.05$ or .1). We consider only the results of the projection estimator, which in the simple setting of the DGP in (20) allows the flexible estimation of the projection without making explicit assumptions about the form of heteroskedasticity. The estimation procedure, choice of approximating functions, and estimation of RMSE for the confidence intervals are as in the section VA.

The results of 200 MC replications are displayed in table 4. For each parameter of interest and for each σ_β, we estimate bias, RMSE, and coverage rates for 95% nominal confidence intervals. We calculate these statistics for a small area where $\beta_a = E(\beta_a) = \tau_{\beta_a,.50}$ and for two areas where the slope is respectively equal to $\tau_{\beta_a,.75}$ and $\tau_{\beta_a,.95}$, where $\tau_{\beta_a,p}$ is the pth quantile of the distribution of β_a. Using the DGP in (20), the true values are calculated as (omitting the subscripts c, h)

$$P(y \leq z | a) = P(\beta_a x + e + \varepsilon \leq z - 10)$$

$$= \Phi\left(\frac{z - 10 - \beta_a 5}{\sqrt{\sigma_e^2 + \sigma_\varepsilon^2 + \beta_a^2}}\right) \quad z = 20, 22$$

$$E(y | a) = 10 + 5\beta_a.$$

Because we are ignoring the presence of heterogeneity in β_a, the prediction for the welfare measure will be the same for all 25 small areas, given that we assume that the distribution of the predictor x is the same across different areas. At the same time, we expect the estimator to perform worse the further away β_a is from its mean value. The results in table 4 are consistent with the intuition. For areas where the slope is equal to the mean value, the bias is negligible and coverage is perfect. When the heterogeneity in β is relatively large (columns 5–8), the coverage is even conservative, because the bias adjustment described in (14) becomes large due to the high intracluster correlation induced by area-specific slopes. The results in panel B show that even when the area-specific slope is equal to the 75th percentile of the distribution, coverage is close to correct, even if, in this case, the bias becomes larger for larger σ_β. For instance, if $\sigma_\beta = .1$, the bias in the calculation of the head count with $z = 20$ is .016, which is approximately 15% of the true value. When we look at areas with slope equal to the 95th percentile, coverage rates are below 75%, and the bias increases further.

Overall, then, heterogeneity in slopes may be as problematic as the presence of area fixed effects for the appli-

TABLE 4.—MONTE CARLO SIMULATIONS: CONSEQUENCES OF AREA HETEROGENEITY IN SLOPES

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	$\sigma_\beta = 0.05$				$\sigma_\beta = 0.1$			
	True	Bias	$\widehat{\text{RMSE}}$	Coverage	True	Bias	$\widehat{\text{RMSE}}$	Coverage
(A)	$\beta_a = 3 = \tau_{\beta,.50}$				$\beta_a = 3 = \tau_{\beta,.50}$			
$P_0(20)$	0.119	0.002	0.023	0.956	0.119	0.003	0.0027	0.960
$P_0(22)$	0.240	0.002	0.029	0.967	0.240	0.006	0.036	0.966
$E(Y)$	25	0.009	0.284	0.978	25	0.003	0.517	1.000
(B)	$\beta_a = 3.03 = \tau_{\beta,.75}$				$\beta_a = 3.07 = \tau_{\beta,.75}$			
$P_0(20)$	0.113	0.008	0.023	0.912	0.107	0.016	0.027	0.891
$P_0(22)$	0.229	0.012	0.029	0.896	0.219	0.027	0.036	0.851
$E(Y)$	25.2	−0.155	0.284	0.885	25.3	−0.326	0.517	0.920
(C)	$\beta_a = 3.12 = \tau_{\beta,.95}$				$\beta_a = 3.23 = \tau_{\beta,.95}$			
$P_0(20)$	0.099	0.017	0.023	0.791	0.081	0.032	0.027	0.634
$P_0(22)$	0.204	0.027	0.029	0.769	0.173	0.056	0.036	0.560
$E(Y)$	25.6	−0.400	0.284	0.582	26.2	−0.815	0.517	0.714

Notes: All results are based on 200 Monte Carlo replications. The true model is $y_{ach} = 10 + \beta_a x_{ach} + e_c + \varepsilon_{ach}$, where $\beta_a \sim N(3, \sigma_\beta^2)$. Coverage rates are calculated for confidence intervals with nominal coverage equal to .95. The RMSE in columns 3 and 7 are the means of the 20 estimated RMSE, *not* the square root of the mean value of the 200 squared deviations of $\hat{W}$ from W. In all simulations, $\sigma_e = .1$, $\sigma_\varepsilon = 3$. Each synthetic sample includes 500 observations, divided into area-specific clusters of 20 households from each small area. $\tau_{\beta,p}$ is the pth quantile of the distribution of β_a. See section VB for further simulation details.

cation of poverty mapping methodologies. Also, in reality, this kind of heterogeneity may take more complex forms, for instance, if the area-specific component of the slopes is correlated with the predictors.

VI. An Empirical Evaluation Using Census Data from Mexico

The MC experiments in section V have demonstrated that even a relatively small amount of heterogeneity in the conditional relation between expenditure and its predictors may lead to severe overstatement of the precision of the resulting estimates. This is our main conclusion. However, it is useful to consider a more concrete example to illustrate our analysis, even though there is no reason to believe that the results will apply everywhere. However, if we find that nominal precision overstates true precision, we know that our general concerns should be taken seriously in practice. In this section, we use census data from Mexico to evaluate the performance of estimators that match census and survey data using the techniques described in the previous sections. The data set is a 10.6% random extract of the 2000 Mexican Census from the Integrated Public Use Micro Sample (IPUMS; Ruggles & Sobek, 1997). Like most other census microdata, the 2000 Mexican Census includes many predictors of income and expenditure, such as housing characteristics, household composition, asset ownership, and occupation and education of each household member. Unlike most other census data sets, this census also includes a measure of individual income during the previous thirty days. This allows us to carry out an experiment that can be summarized as follows. First, we identify relatively large regions (the states of Chiapas, Oaxaca, and Veracruz), from which we select a synthetic household survey by drawing a random sample of household-specific observations of in-

come (y) and a set of predictors (x). We use this sample to estimate the parameters of a model for the probability of being poor (of income being below a fixed poverty line z) conditional on a set of predictors. We then merge these parameters with census information on the predictors for the whole population in the region. This allows us to calculate point estimates and standard errors of predictions of income-based poverty measures defined for a list of small areas within the same region. While keeping the census populations constant, we repeat the synthetic survey sample generation and the two-step estimation procedure a large number of times. For each small area, we calculate coverage rates of nominal 95% confidence intervals as the fraction of times that the true value of the poverty measure lies within the interval. If the conditional model in each small area is the same as in the larger region, coverage rates should be approximately equal to the nominal rates.[10] If instead, coverage rates are much lower than .95, substantial heterogeneity is likely to exist, and the variance of estimators based on conditional independence assumptions will underestimate the true variance of the prediction error. In Demombynes et al. (2007), a similar exercise is completed to evaluate the performance of ELL by using data from a complete census from Mexican areas where the well-known welfare program PROGRESA has been implemented. However, Demombynes et al. (2007) use small areas generated by aggregating villages that are selected at random, and hence impose by construction the approximate validity of area homogeneity; in consequence, their results are not likely to be informative about the effects of heterogeneity, which has been removed by construction. In our case, areas

[10] As described in section VIA, we assume that the true values are identical to the estimates obtained from the census extract.

FIGURE 1.—MAP OF *MUNICIPIOS,* STATE OF CHIAPAS, MEXICO
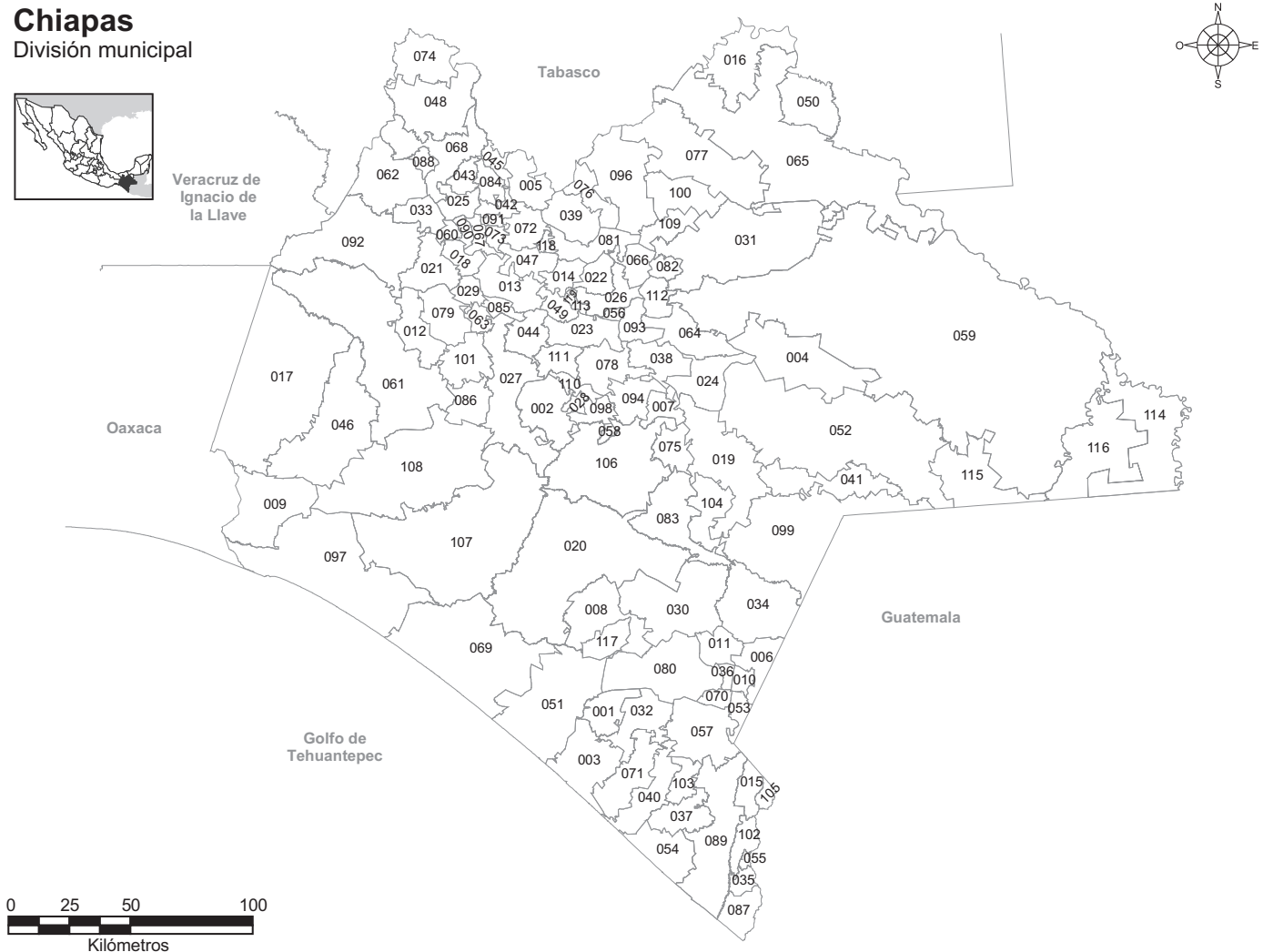
Source: INEGI, Mexico.
Note: The map illustrates the 119 *municipios* that form the state of Chiapas according to the 2005 geostatistical census of Mexico.

coincide with actual administrative units (*municipios*), so that the results of the empirical validation will show the consequences of the failure of homogeneity for poverty estimates in actual *municipios*. A validation exercise similar to the one presented here has been developed independently by Elbers, Lanjouw, and Leite (2008), who find that in the context of the Brazilian state of Minas Gerais, ELL performs well.

The details of the validation experiment are described in section VIA; readers interested only in the results can refer directly to section VIB.

A. Details of the Validation Exercise

The complete IPUMS microdata extract for Mexico 2000 contains more than 10 million observations, so to keep the validation exercise manageable, we limit our analysis to the rural section of three of the largest Mexican states: Chiapas, Oaxaca, and Veracruz. Each state is subdivided into a large

number of *municipios,* and we treat each state as a separate region, and each *municipio* as a small area. The map in figure 1 shows the subdivision of the state of Chiapas into *municipios* according to the 2005 geostatistical census of Mexico.[11] Clearly most areas are very small, and in practical applications, household survey data alone would not be sufficient to estimate welfare measures with acceptable precision for areas smaller than a state. For instance, the state-specific rural sample size in the 2002 Mexican Family Life Survey (MFLS) ranged from 47 (Distrito Federal) to 469 (in Michoacán).[12] Most *municipios* are instead not represented at all in the rural sample, and of 73 *municipios* included in the survey sample, only two have more than 54 observations. The actual (census) population size of *muni-*

[11] See http://www.cuentame.inegi.gob.mx. The subdivisions in 2000 and 2005 were essentially identical.
[12] For these calculations, we have classified households as rural when they live in communities with population below 2,500.

TABLE 5.—MEXICO 2000 PSEUDO-CENSUS: SUMMARY STATISTICS

	Chiapas	Oaxaca	Veracruz
Extract size (number of households)	58,358	120,934	96,826
Pseudo-census household population size	398,347	402,098	621,609
Pseudo-census individual population size	2,052,071	1,897,684	2,834,599
No. of *municipios*	118	562	210
Mean number of households in a pseudo-census *municipio*	3,375	715	2,960
Median number of households in a pseudo-census *municipio*	2,790	423	2,028
Fraction of households reporting zero income	.185	.223	.142
Fraction of individuals with missing income	.018	.017	.016
Mean monthly income per head (2000 pesos)	487	297	407
Median monthly income per head (2000 pesos)	99	90	235
Poverty head count ratios (line 200 pesos day/person)			
Mean	.67	.64	.43
Median	.70	.69	.42
Standard deviation	.18	.21	.19
5th percentile	.33	.21	.14
95th percentile	.90	.92	.76

Source: Authors' calculations from Mexico 2000 Census IPUMS extract (rural only).
Note: See section IVA for details about the construction of the pseudo-census. The statistics for the head count ratios refer to the distribution of the *municipio*-specific ratios in each state.

cipios is very heterogeneous but relatively large. The median household population size of the rural sector of a *municipio* is 2,790 in Chiapas, 423 in Oaxaca, and 2,028 in Veracruz (see table 5). Hence, our choice of using *municipios* as small areas.

Because we wish to work with a census, while IPUMS includes only a 10.6% extract of the complete microdata, we first generate a complete pseudo-census with a number of observations equal to actual census population. For this purpose, we generate a pseudo-census of size analogous to the complete Census 2000 by expanding the extract. This is done by replacing each observation in the extract with identical replicates in number identical to the (integer) weight provided in the data set. The pseudo-census so created is then treated as the actual (nonrandom) population of interest. Because the census extract does not include identifiers for separate census EAs, we cannot include in the analysis cluster means of household-level variables measured in the census. ELL suggest this strategy to reduce the extent of intracluster correlation in the data. As an alternative, we include among the predictors census means of household-level variables calculated for each *municipio*.

We assume that the object of interest is a poverty map for all *municipios* in the three Mexican states, but the researcher has access only to a (pseudo) household survey, defined here as ten observations from each of fifty *municipios* selected at random without replacement. By construction, this sampling design leads to a different probability of selection for different households, so that estimation is done after constructing sampling weights. We classify a household as poor if total monthly income per head y is below a threshold z equal to 200 pesos.[13] Because the census includes income for all households, the true value of the head count ratio for each *municipio* in the expanded extract can be calculated as the proportion of individuals who live in households with

per capita income below z. These true head count ratios can then be compared with the corresponding estimates obtained using synthetic survey samples and making use of a (possibly incorrect) conditional independence assumption.

In each of the three states, there is considerable heterogeneity in the distribution of the *municipio*-specific poverty head counts. Table 5 shows that Veracruz is the least poor state, with a median head count equal to .41; both Chiapas and Oaxaca are poorer, with median poverty rates close to 70%.

We evaluate the coverage of 95% nominal confidence intervals with 250 MC simulations. We complete independent simulations for the three states of Chiapas, Oaxaca, and Veracruz. Throughout the simulations, we treat the pseudo-census generated as described in the previous paragraph as the true population, and in each replication we draw a different random sample without replacement from such population. Because each sample is represented by a subset of the census data, assumption MP holds by construction. Table 6 provides a list of the predictors in the first stage.[14] Once the artificial sample has been selected, we calculate point estimates and the corresponding RMSEs for each *municipio* using the projection estimator described in section IVA. Finally, we record if the true value of the head count ratio in each *municipio* lies within the interval boundaries.

A few caveats should be kept in mind. First, we reiterate that we do not have access to a full census, only to an extract. Ideally, a validation would proceed by drawing survey samples from a true census and verifying over a large number of replications whether the true values calculated

[13] The USD Mexican peso PPP exchange rate in 2000 was 6.79, so that 200 pesos corresponds to approximately 1 PPP dollar per person per day (Heston, Summers, & Aten, 2006).

[14] Including a large number of regressors may lead to overfitting. We have attempted an alternative procedure where, for each sample, the set of predictors is chosen using the following criterion. First, regressors are sorted according to the pseudo-R^2 of univariate logit regressions. Then we determine the set of the first k regressors to include in the model, where k maximizes a Bayesian information criterion (Schwarz, 1978). This alternative procedure worsens coverage considerably, so we do not include the results here.

TABLE 6.—MEXICO 2000 PSEUDO-CENSUS: VARIABLES USED AS PREDICTORS

Head is literate
Access to electricity
Owns refrigerator
Owns TV
Owns radio
Number of rooms
Access to toilet within dwelling
Age of head
Head belongs to indigenous group
Main cooking fuel is wood
Dwelling has dirt floor
Primary dwelling material is brick or stone
Primary roof material is masonry, concrete, or tile
Speaks only indigenous language
Speaks both indigenous language and Spanish
Head is working
Head works in agriculture/fishery/forestry/mining
Number of household members ages 0–12 (and its squared)
Number of household members older than 65 (and its squared)
Number of male members ages 13–65 (and its squared)
Number of female members age 13–65 (and its squared)
Head is a woman
Municipio-level means:
    Head is literate
    Years of schooling of head
    Access to electricity
    Owns radio
    Access to toilet within dwelling
    Dwelling has dirt floor
    Primary dwelling material is brick or stone
    Primary roof material is masonry, concrete, or tile
    Speaks only indigenous language
    Head works in agriculture/fishery/forestry/mining

Source: IPUMS Mexico Census 2000.
Note: List of variables used as predictors for a binary variable equal to 1 if household monthly income per head is below the poverty line.

from the census lie within the 95% confidence intervals. Our choice of working with an expanded extract is a way to recreate a framework close to this ideal, given the data constraints. Second, the census extract does not report identifiers for census enumeration areas. Hence, in implementing the projection estimator, we estimate the covariance term in the bias correction in (14) treating *municipios* as clusters. Third, the income measures included in the census may not be as accurate as the income or expenditure measures assessed in household surveys where the measurement of living standards is often the main objective. For instance, a nonnegligible fraction of households report zero income over the previous 30 days (see table 5). However, a comparison between census 2000 and the 2002 MFLS does not show major discrepancies. In rural Oaxaca, median monthly per capita income in 2002 pesos was 80.8 pesos according to MFLS, and 100 pesos according to census 2000. In the state of Veracruz, the MFLS median was 233.3 pesos, while the census estimate was 263.[15]

[15] The consumer price index increased from 100 to 111.7 between 2000 and 2002 (data extracted on February 20, 2008, from http://stats.oecd.org/wbos). We do not report on a comparison for Chiapas because this state was not separately identified in MFLS 2002.

B.   Results of the Validation Exercise

The predictors listed in table 6 have moderately good explanatory power in each of the three states. When all of the complete observations from the census are used, the pseudo-R^2 is .1886 in Chiapas, .1684 in Oaxaca, and .1652 for Veracruz.[16] For each state, we display in figure 2 histograms of coverage rates. Each observation shows, for a given *municipio,* the fraction of MC replications for which the true value of the poverty head count ratio lies within a nominal 95 percent confidence interval. Because assumption MP holds by construction, deviation of the coverage rates from the nominal ones likely indicates failure of the homogeneity assumption AH.

The histograms in figure 2 show that while coverage rates for most areas are not far from the nominal 95%, in a large fraction of areas, coverage rates are well below the nominal level. This indicates the existence of heterogeneity across *municipios,* which an estimator that relies on the area homogeneity assumption AH ignores. The fraction of *municipios* where coverage remains below .75 is .33 in Chiapas, .50 in Oaxaca, and .48 in Veracruz. In all three states, coverage rates are below 50% in approximately 10% of *municipios.* Although the estimated confidence intervals appear to systematically overstate the precision of the estimator, they are relatively wide. The mean width of a confidence interval is .33 in Chiapas (minimum .19 and maximum .60), .41 in Oaxaca (minimum .23 and maximum 1.92), and .36 in Veracruz (minimum .20 and maximum .83). It should be noted that poor coverage is not a product of our area sizes being smaller than the ones that would typically be used in poverty mapping. First, poor coverage rates do not arise only in the smallest areas. Second, MC results show that when—as in our validation exercise—the size of the survey sample is not very large, confidence intervals have actually better coverage for small areas (see Appendix B). So there is no reason to suppose that coverage would be better if we had chosen larger areas.

These findings suggest that heterogeneity in the conditional distribution of income given the predictors is a condition that may arise in empirical settings and is not just a complication of theoretical interest. Of course, the results discussed in this section do not imply that similar heterogeneity will be present elsewhere (any more than would the absence of heterogeneity have meant that it is absent everywhere), although the plausibility of spatial heterogeneity in intercepts or rates of return suggests that at the least, it would be unwise to assume it away. Indeed, even in the context of this empirical exercise, we find a certain degree of variation in the distribution of coverage rates across *municipios* among the three different states. Specifically, Oaxaca is the state where the distribution appears to be more skewed to the left, that is, with low coverage rates for a larger fraction of *municipios.*

[16] The full estimation results are available on request.

FIGURE 2.—DISTRIBUTION OF COVERAGE RATES BY STATE, POVERTY HEAD COUNTS
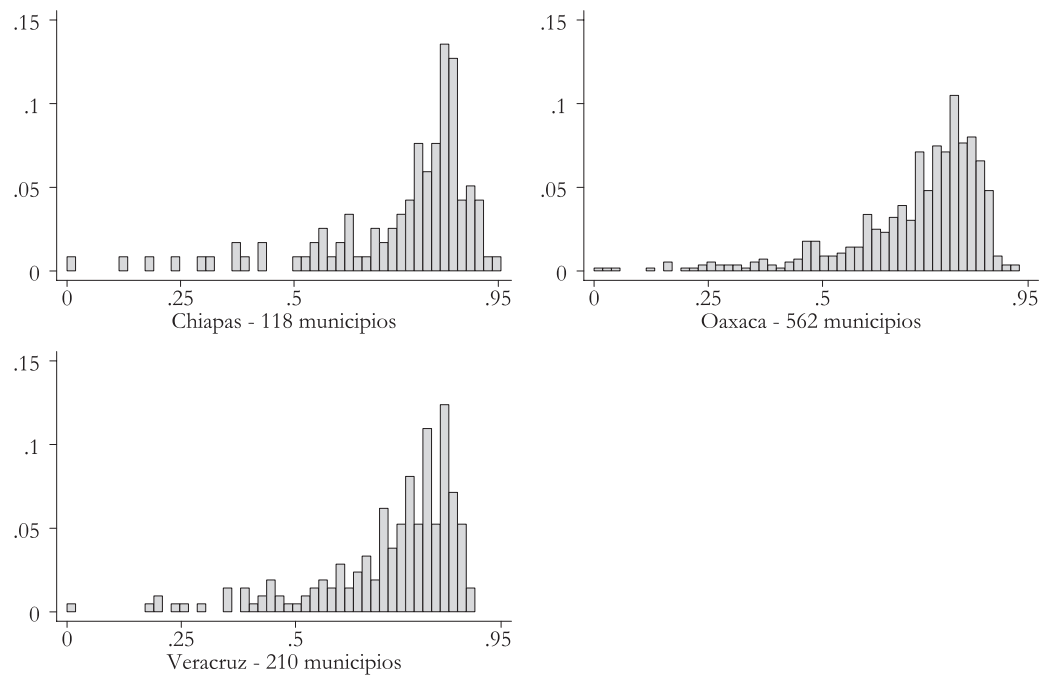
Source: Authors' calculations from IPUMS Mexico 2000 Census extract.

Note: An observation indicates, for a given *municipio,* the fraction of Monte Carlo replications for which the true value of the poverty head count ratio lies within a nominal 95% confidence interval.

VII.   Conclusion

Large household expenditure survey data are not suited for the construction of precise welfare estimates for small areas, because at most, a handful of observations are usually available from geographical entities of limited size. However, recent years have seen an increasing availability of "poverty maps" for small areas in developing countries. These maps are usually constructed using a methodology developed in Elbers et al. (2003), which exploits the possibility of merging data from a census and a household survey to improve precision of estimates for small areas. This methodology is deemed able to allow the estimation of welfare estimates for areas of fewer than 20,000 households as precise as those otherwise obtainable with survey data alone only for populations hundreds of times larger. In this paper, we argue, first, that there is no general reason to suppose that the conditions that are necessary to match survey and census data will hold in practice. Second, we argue that estimates based on those assumptions may severely underestimate the variance of the error in predicting welfare estimates at the local level (and hence severely overstate the coverage of confidence intervals) in the likely presence of small-area heterogeneity in the conditional distribution of expenditure or income. The presence of area heterogeneity is apparent in an empirical experiment carried out with data from the 2000 Mexican Census.

This experiment shows that our theoretical concerns can be important in real examples, though we do not argue that they will be so in every case, as the results in Elbers et al.

(2008) show. Overall, we believe that efforts to calculate welfare estimates for small areas merging survey and census data are worthwhile, but we also believe that the current literature has not sufficiently emphasized the limitations of the current methodologies and the strong assumptions that they require in order to permit meaningful inference. Such limitations should be stressed, and the precision of the estimates should be judged accordingly. Policymakers who make use of poverty maps to allocate funds and improve targeting the welfare programs should be aware that such maps may be subject to more uncertainty and error than has been claimed in the literature on poverty mapping and take into account the misallocation of funds that will follow the misidentification of their targets.

REFERENCES

Alderman, Harold, Miriam Babita, Gabriel Demombynes, Ntabiseng Ma-khatha, and Berk Özler, "How Low Can You Go? Combining Census and Survey Data for Mapping Poverty in South Africa," *Journal of African Economies* 11:3 (2003), 169–200.

Chen, Shaohua, and Martin Ravallion, "How Have the World's Poorest Fared since the Early 1980s?" *World Bank Research Observer* 19:2 (2004), 141–169.

Chen, Xiaohong, Han Hong, and Elie Tamer, "Measurement Error Models with Auxiliary Data," *Review of Economic Studies* 72:2 (2005), 343–366.

Chen, Xiaohong, Han Hong, and Alessandro Tarozzi, "Semiparametric Efficiency in GMM Models with Auxiliary Data," *Annals of Statistics* 36:2 (2008), 808–843.

Citro, Constance F., and Graham Kalton (Eds.), *Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond* (Washington, DC: National Academy Press, 2000).

Cochrane, William G., *Sampling Techniques* (New York: Wiley, 1977).

Deaton, Angus, and Margaret Grosh, "Consumption" (pp. 91–133), in Margaret Grosh and Paul Glewwe (Eds.), *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study* (New York: Oxford University Press for the World Bank, 2000).

Demombynes, Gabriel, Chris Elbers, Jean O. Lanjouw, and Peter Lanjouw, "How Good a Map? Putting Small Area Estimation to the Test," World Bank policy research working paper no. 4155 (2007).

Demombynes, Gabriel, and Berk Özler, "Crime and Local Inequality in South Africa," *Journal of Development Economics* 76:2 (2005), 265–292.

Dorfman, Robert, "A Formula for the Gini Coefficient," this REVIEW 61:1 (1979), 146–149.

Elbers, Chris, Tomoki Fujii, Peter Lanjouw, Berk Özler, and Wesley Yin, "Poverty Alleviation through Geographic Targeting: How Much Does Disaggregation Help?" *Journal of Development Economics* 83 (2007), 198–213.

Elbers, Chris, Peter Lanjouw, Johan Mistiaen, Berk Özler, and Ken Simler, "On the Unequal Inequality of Poor Communities," *World Bank Economic Review* 18:3 (2004), 401–421.

Elbers, Chris, Jean O. Lanjouw, and Peter Lanjouw, "Micro-Level Estimation of Welfare," World Bank policy research working paper no. 2911 (2002).

Elbers, C., J. Lanjouw, and P. Lanjouw, "Micro-Level Estimation of Poverty and Inequality," *Econometrica* 71:1 (2003), 355–364.

Elbers, Chris, Peter Lanjouw, and Philippe George Leite, "Brazil within Brazil: Testing the Poverty Map Methodology in Minas Gerais," World Bank policy research working paper no. 4513 (2008).

Grosh, M., and J. N. K. Rao, "Small Area Estimation: An Appraisal," *Statistical Science* 9:1 (1994), 55–93.

Heckman, J., R. LaLonde, and J. Smith, "The Economics and Econometrics of Active Labor Market Programs" (pp. 1865–2097), in O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics* (Amsterdam: Elsevier Science, 1999).

Heston, Alan, Robert Summers, and Bettina Aten, "Penn World Table Version 6.2" (Philadelphia Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania, 2006). http://pwt.econ.upenn.edu/php_site/pwt_index.php.

Imbens, Guido, "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," this REVIEW 86:1 (2004), 4–29.

Kish, Leslie, *Survey Sampling* (New York: Wiley, 1965).

Lee, L., and J. Sepanski, "Estimation of Linear and Nonlinear Errors-in-Variables Models Using Validation Data," *Journal of the American Statistical Association* 90:429 (1995), 130–140.

Little, Roderick, and Donald Rubin, *Statistical Analysis with Missing Data,* 2nd ed. (New York: Wiley, 2002).

Mistiaen, J. A., B. Özler, T. Razafimanantena, and J. Razafindravonona, "Putting Welfare on the Map in Madagascar," World Bank, African Region working paper no. 34 (2002).

National Research Council, *Panel on Small-Area Estimates of Population and Income: Estimating Population and Income of Small Areas* (Washington, DC: National Academy Press, 1980).

Rosenbaum, Paul R., and Donald B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70:1 (1983), 41–55.

Rubin, Donald, "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies," *Journal of Educational Psychology* 66 (1974), 688–701.

———— "Inference and Missing Data," *Biometrika* 63 (1976), 581–592.

———— *Multiple Imputation for Nonresponse in Surveys* (New York: Wiley, 1987).

Ruggles, S., and M. Sobek, "Integrated Public Use Microdata Series: Version 2.0" (Minneapolis: Historical Census Projects, University of Minnesota, 1997). http://www.ipums.umn.edu.

Schwarz, Gideon, "Estimating the Dimension of a Model," *Annals of Statistics* 6:2 (1978), 461–464.

Wooldridge, Jeffrey, *Econometrics of Cross Section and Panel Data* (Cambridge, MA: MIT Press, 2002a).

———— "Inverse Probability Weighted M-Estimators for Sample Selection, Attrition and Stratification," *Portuguese Economic Journal* 1 (2002b), 117–139.

APPENDIX A:

Proofs

Proof of equation (5).

$$
\begin{aligned}
0 &= E[s_h g(y_h; W_0) \mid h \in H(A)] \\
&= E\{E[s_h g(y_h; W_0) \mid X_h, h \in H(A)]\} \\
&= \int_X E[s_h g(y_h; W_0) \mid X_h, h \in H(A)] dF(X_h \mid h \in H(A)) \\
&= \int_X E[s_h g(y_h; W_0) \mid X_h, h \in H(R)] dF(X_h \mid h \in H(A)),
\end{aligned}
$$

where the last step follows from AH, and MP guarantees that the correlates in the census and in the survey are measured in the same way.

Proof of equation (8).

$$
\begin{aligned}
\mathrm{Var}\,(\hat{\mu}_y) &= \mathrm{Var}\,(\mu_y - \hat{\mu}_y) = \left(\frac{1}{\sum_{c=1}^C m_c}\right)^2 \mathrm{Var}\left(\sum_{c=1}^C \sum_{h=1}^{m_c} u_{ch}\right) \\
&= \left(\frac{1}{\sum_{c=1}^C m_c}\right)^2 \left\{ \sum_{c=1}^C \sum_{h=1}^{m_c} \sigma_u^2 + \sum_{c=1}^C \sum_{h=1}^{m_c} \left(\sum_{c'=1, c'\neq c}^C \sum_{h'=1}^{m_{c'}} \rho_a \sigma_u^2 \right) \right. \\
&\quad \left. + \sum_{c=1}^C \sum_{h=1}^{m_c} \left(\sum_{h'=1, h'\neq h}^{m_c} \rho_c \sigma_u^2 \right) \right\} \\
&= \left(\frac{\sigma_u}{\sum_{c=1}^C m_c}\right)^2 \left\{ \sum_{c=1}^C m_c + \sum_{c=1}^C \sum_{h=1}^{m_c} \left(\sum_{c'=1, c'\neq c}^C m_{c'}\rho_a \right) \right. \\
&\quad \left. + \sum_{c=1}^C m_c(m_c - 1)\rho_c \right\} \\
&= \left(\frac{\sigma_u}{\sum_{c=1}^C m_c}\right)^2 \\
&\quad \times \left\{ \sum_{c=1}^C m_c + \sum_{c=1}^C \sum_{c'=1, c'\neq c}^C m_c m_{c'}\rho_a + \sum_{c=1}^C m_c(m_c - 1)\rho_c \right\}.
\end{aligned}
$$

APPENDIX B:

Bias Correction

In section IVA, we explained that the parametric projection estimator in (12) would in general be different from the true head count (11) even if the first-stage parameters γ were known. The difference between the two quantities, which we refer to as bias, must be taken into account in the construction of the MSE. In this appendix, we derive an expression for the bias that can be estimated using the survey data, and we describe an MC experiment to evaluate the performance of the estimator under a variety of conditions.

Let $p_h(\gamma) \equiv P(y_h \leq z \mid X_h; \gamma)$, and let p_h denote the value of $p_h(\gamma)$ evaluated at the true value of the parameters. By definition, the bias for a given area A is

TABLE 7.—COVERAGE OF PROJECTION ESTIMATOR

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	$\rho = 0$		$\rho = .02$		$\rho = .05$		$\rho = .10$	
	s.e. Only	Bias Correction	s.e. Only	Bias Correction	s.e. Only	Bias Correction	s.e. Only	Bias Correction
10 Households from 50 Areas								
n = 100								
$R^2 = .20$	0.62	0.98	0.48	0.93	0.38	0.89	0.31	0.92
$R^2 = .60$	0.60	0.96	0.55	0.95	0.43	0.90	0.42	0.90
n = 500								
$R^2 = .20$	0.83	0.97	0.56	0.85	0.43	0.84	0.31	0.88
$R^2 = .60$	0.83	0.96	0.63	0.88	0.52	0.84	0.44	0.86
n = 5,000								
$R^2 = .20$	0.93	0.96	0.61	0.80	0.47	0.80	0.32	0.87
$R^2 = .60$	0.93	0.96	0.65	0.84	0.52	0.80	0.41	0.84
n = 15,000								
$R^2 = .20$	0.94	0.95	0.59	0.80	0.45	0.81	0.35	0.87
$R^2 = .60$	0.95	0.96	0.68	0.82	0.53	0.82	0.40	0.82
20 Households from 1,000 Areas								
n = 100								
$R^2 = .20$	0.13	0.96	0.08	0.94	0.06	0.95	0.05	0.95
$R^2 = .60$	0.12	0.94	0.08	0.94	0.09	0.95	0.08	0.96
n = 500								
$R^2 = .20$	0.24	0.97	0.11	0.94	0.08	0.95	0.06	0.95
$R^2 = .60$	0.23	0.98	0.13	0.94	0.10	0.96	0.08	0.95
n = 5000								
$R^2 = .20$	0.61	0.98	0.12	0.93	0.08	0.96	0.06	0.96
$R^2 = .60$	0.63	0.98	0.14	0.93	0.10	0.95	0.06	0.94
n = 15,000								
$R^2 = .20$	0.81	0.98	0.12	0.94	0.08	0.95	0.06	0.95
$R^2 = .60$	0.80	0.97	0.12	0.92	0.09	0.96	0.08	0.94

Notes: Figures are coverage rates for 95% confidence intervals, calculated over 1,000 Monte Carlo replications. Confidence intervals are calculated either taking into account only the standard error of the estimator (s.e. only) or including an estimate of the bias squared (Bias Correction). The DGP is $y_{ah} = 20 + \beta x_{ah} + \eta_a + \varepsilon_{ah}$, $x \sim N(5, 1)$, $\varepsilon \sim N(0, 1)$, $\eta \sim N(0, \sigma_\eta^2)$. In each replication, the estimated parameter is $P(y_{ah} \leq z)$, where z is the 25th percentile of the distribution of y when the area fixed effect η is 0. In each cell, the parameters β and σ_η^2 are uniquely determined by the values of ρ and R^2 relevant for that cell (see Appendix B for details). The parameter n indicates the size of the small area population, which is generated in each replication as a random draw from the DGP.

$$bias(\hat{W}) = E\left[\frac{1}{N_A} \sum_{h \in H(A)} p_h(\hat{\gamma})\right] - \frac{1}{N_A} \sum_{h \in H(A)} 1(y_h \leq z)$$

$$= \frac{1}{N_A} \sum_{h \in H(A)} [E(p_h(\hat{\gamma})) - 1(y_h \leq z)].$$

This quantity is unknown and depends on the specific small area being considered. If we approximate $E(p_h(\hat{\gamma}))$ with its true value p_h, we have

$$bias^2(\hat{W}) \approx \left[\frac{1}{N_A} \sum_{h \in H(A)} (p_h - 1(y_h \leq z))\right]^2$$

$$= \frac{1}{N_A^2} \sum_{h \in H(A)} \sum_{h' \in H(A)} [p_h - 1(y_h \leq z)][p_{h'} - 1(y_{h'} \leq z)]$$

$$= \frac{1}{N_A^2} \sum_{h \in H(A)} [p_h - 1(y_h \leq z)]^2 \qquad (21)$$

$$+ \frac{1}{N_A^2} \sum_{h \in H(A)} \sum_{h' \in H(A), h' \neq h} [p_h - 1(y_h \leq z)][p_{h'} - 1(y_{h'} \leq z)].$$

Both terms on the right-hand side of (21) include the value of y_h for all observations in the area and are therefore unknown. Expression (14) follows from simple manipulation after replacing the elements in the summations with sample estimates of their expected value. Note that while the first term in (14) goes to 0 when the size of the area increases, the second term does not.

B.1 Monte Carlo Experiments

Assume that expenditure can be modeled as:

$$y_{ah} = 20 + \beta X_{ah} + \eta_a + \varepsilon_{ah},$$

$$X_{ah} \sim N(5, 1), \quad \varepsilon_{ah} \sim N(0, 1), \quad \eta_a \sim N(0, \sigma_\eta^2),$$

where X, η, and ε are independent. The DGP mimics the framework relevant for the pseudo-validation exercise with Mexican data in section VI. We experiment with different models where we vary both the size of the small area and the magnitude of the R^2 and the intracluster correlation coefficient $\rho = \sigma_\eta^2/(\sigma_\eta^2 + \sigma_\varepsilon^2)$. Given that σ_ε^2 is kept equal to 1, the choice of ρ uniquely determines the value of σ_η^2. Finally, the choice of ρ and R^2 uniquely determines the value of β, which completes the DGP. This last result follows noting that $R^2 = 1 - \text{var}(\eta_a + \varepsilon_{ah})/\text{var}(y_{ah})$ and $\beta = \sqrt{R^2(1 + \sigma_\eta^2)/(1 - R^2)}$.

We experiment with values of R^2 and ρ in a range consistent with the poverty mapping literature. We set $R^2 \in \{.20, .60\}$ and $\rho \in \{0, .02, .05, .10\}$. For each combination, small areas have size $N_A \in \{100, 500, 5000, 15000\}$. Each coverage rate is calculated over 1,000 MC replications. In each replication, we first generate a small area drawing a single η_a from its distribution, and then we generate a synthetic sample drawing either 10 units from 50 different areas (that is, drawing a different η_a for each area) or 20 units from 1,000 areas. Hence, the first-stage estimation is implemented with either 500 or 20,000 observations. In all simulations, the poverty line z is set at a value equal to the 25th percentile of the distribution of y when the area fixed effect η_a is 0, that is, z solves $P(20 + \beta x + \varepsilon < z) = .25$. The assumed DGP implies that $z = 20 + 5\beta + \Phi^{-1}(.25)\sqrt{\beta^2 + 1}$, where Φ^{-1} indicates the inverse of the cumulative distribution function of a standard normal. We show the results in table 7. It is apparent that if the bias is not taken into account (columns 1, 3, 5, and

7), the projection estimator systematically underestimates the true prediction error, so that coverage rates are almost always below .95, often by a substantial amount. Even with no location effects (column 1), coverage rates are not correct unless the area includes a large number of units. As expected, coverage rates throughout the table are lower when the intracluster correlation is large. Increases in population size do not lead to systematic improvements in coverage, because all observations within the same area share the same fixed effect η, which therefore does not average out. Note also that coverage worsens when the synthetic sample becomes larger. This is because confidence intervals in columns 1, 3, 5, and 7 are calculated taking into account only the component of the MSE that derives from estimation error, so that the fraction of the MSE accounted for by the bias (and disregarded in the calculation of the confidence interval) becomes larger moving from the top to the bottom panel of the table. The results in columns 2, 4, 6, and 8 show that coverage rates improve dramatically when the bias correction is taken into account. When $n = 100$, coverage rates are always almost identical to the nominal ones. This is also always true when the synthetic sample is large. When the sample is small (top panel), $\rho > 0$, and population size is large (500 or above), coverage rates remain in the .80 to .88 range, so that the bias adjustment systematically understates the MSE, even if not by much. This is likely due to the fact that when the sample size is small, the calculation of the covariance term in (14) sometimes results in a negative number even if the covariance is actually positive. This has the effect of reducing the estimated RMSE, even if the covariance should contribute to its increase. Indeed, the results in the bottom panel show that coverage rates are essentially identical to nominal ones when the synthetic sample becomes large, in which case the covariance can be estimated precisely.