Using Census and Survey Data to Estimate Poverty and Inequality for Small Areas

Alessandro Tarozzi
Duke University

Angus Deaton
Princeton University

June 2007*

Abstract

Household expenditure survey data cannot yield precise estimates of poverty or inequality for small areas for which no or few observations are available. Census data are more plentiful, but typically exclude income and expenditure data. Recent years have seen a widespread use of small-area "poverty maps" based on census data enriched by relationships estimated from household surveys that predict variables not covered by the census. These methods are used to estimate putatively precise estimates of poverty and inequality for areas as small as 20,000 households. In this paper 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. The relationships that are used to bridge the surveys and censuses are not structural but are projections of missing variables on a subset of those variables that happen to be common to the survey and the census supplemented by local census means appended to the survey. As such, the coefficients of the projections will generally vary from area to area in response to variables that are not included in the analysis. Estimates of poverty and inequality that assume homogeneity will generally be inconsistent in the presence of spatial heterogeneity, and error variances calculated on the assumption of homogeneity will underestimate mean squared errors and overestimate the coverage of calculated confidence intervals. We use data from the 2000 census of Mexico to construct synthetic "household surveys" and to simulate the poverty mapping process using a robust method of estimation; our simulations show that while the poverty maps contain useful information, their nominal confidence intervals give a misleading idea of precision. **JEL: I32, C31, C42**

Key words: Small Area Statistics, Poverty, Inequality, Heterogeneity, Survey Methods.

*We thank Gabriel Demombynes, Chris Elbers, Han Hong, Shakeeb Khan, Peter Lanjouw, Phillippe Leite, Barbara Rossi and seminar participants at the 2006 North Eastern Universities Development Conference (Cornell) and the University of Mannheim for useful comments and conversations, and IPUMS for access to a 2000 Mexican Census Extract. Maria Eugenia Genoni provided excellent research assistance. We are solely responsible for all errors and omissions. Alessandro Tarozzi, Dept of Economics, Duke University, Social Sciences Building, PO Box 90097, Durham, NC 27708, taroz@econ.duke.edu. Angus Deaton, 328 Wallace Hall, Woodrow Wilson School, Princeton University, Princeton, NJ 08544

1 Introduction

Household surveys collect information on incomes, expenditures, and demographics, and are regularly used to generate population statistics, such as mean incomes, poverty headcount 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 and Ravallion (2004). Some of these surveys support sub-national 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 obtains good income information for five percent 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 and Rao (1994), Citro and Kalton (2000)).

In most developing countries, censuses do not collect income or expenditure information, so that small area poverty estimates are typically not available even in 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 substantial 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 ten percent 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, of the effects of inequality on crime, of whether inequality is higher among the poor, and of 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, nor a consideration of the likelihood or consequences of assumption failure.

In this paper, we provide a set of "conditional independence" or "area homogeneity" assump-

[1] For a comprehensive description of the methodology used by the Bank, as well as for reference to the numerous applications, see www.worldbank.org/poverty.

[2] See www.worldbank.org/research/povertymaps/kenya.

tions 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 of a more robust 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 that their coefficients will generally be functions of any local variables that are not explicitly included.

We consider some obvious special cases of heterogeneity—for example, where there is a small area random effect that makes everyone in the area better or worse-off than would be predicted by the projection—and discuss the consequences for estimators, focusing 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 seriously 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. Since the Mexican census contains income information, these can be checked against the true poverty rates for small areas. 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 3 describes the consequences of unobserved heterogeneity across areas. Section 4 describes estimation, and proposes a nonparametric estimator that is more misspecification-robust, simpler and computationally faster than the estimator in Elbers et al. 2003. Section 5 describes a series of Monte Carlo simulations that compare the performance of the two estimators in a variety of contexts. Section 6 describes the validation exercise with data from the 2000 Mexican Census. Section 7 concludes.

2 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 which are 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 (as well as some inequality measures) are identified by a simple population moment condition such as the following:

$$E\left[s_h g(y_h; W_0) \mid h \in H(A)\right] = 0, \tag{1}$$

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

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

where $\alpha \geq 0$ is a known parameter and $1(E)$ is an indicator equal to one when event E is true. When $\alpha = 0$ the index becomes the headcount 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 of them 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 $E[y \ln y]/E(y) - \ln(E(y))$, while the Atkinson inequality index is

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

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

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

Before proceeding further, it is useful to describe the kind of data on which estimation is based. There are two data sources. The first is a household survey of region R with 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 of which is 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 analogue of the corresponding moment condition. As an example, the FGT poverty index can be estimated as

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

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, if these are small enough. Sample size would instead be more than adequate in a complete census of the small area, which will typically include several thousands of households. Censuses, though, rarely include reliable information on income or expenditure. However, as a rule 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 ***(CI) Conditional Independence (or Area Homogeneity):***

$$f\left(y_h \mid X_h, h \in H(A)\right) = f\left(y_h \mid X_h, h \in H(R)\right). \tag{4}$$

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 non-nested) coding schemes for schooling, industry or occupation of household members. Different reports may also arise from other less obvious reasons, even if census and household survey use the exact 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 CI 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 CI 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 et al. (1999)). 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 and Rubin (2002), Wooldridge (2002b)). Analogous assumptions can be found in the estimation of non-linear models with non-classical measurement errors in presence of validation data. In this case the requirement is that the distribution of the mis-measured variables conditional on a set of proxies is the same in the main and in the auxiliary sample (see e.g. Lee and Sepanski (1995), Chen et al. (2005), Chen et al. (2007)).

In the estimation of small area statistics, Assumption CI 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 are 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 CI; 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 that caution should be exercised

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

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

A weaker (but still restrictive) assumption is the following:

(MCI) Mean Conditional Independence:

$$E\left[s_h g(y_h; W_0) \mid X, h \in H(A)\right] = E\left[s_h g(y_h; W_0) \mid X, h \in H(R)\right]. \tag{5}$$

Assumptions such as CI or MCI require, 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 MCI hold, the welfare estimate of interest is also identified by the following modified moment condition:

$$\int E\left[s_h g(y_h; W_0) \mid X, h \in H(R)\right] dF(X \mid h \in H(A)) = 0, \tag{6}$$

where $dF(.)$ represents the distribution of the correlates in the small area.[4] In Appendix A we show that (6) can be obtained from assumptions MP and MCI from a simple manipulation of the moment condition (1). If we replace the modified moment condition (6) 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.

3 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 CI. Virtually all household expenditure surveys adopt a complex survey design, so that enumeration areas (EAs) such as villages or urban blocks are sampled first, and then households are sampled from each EA. As is well known, the resulting intra-cluster correlation among households drawn from the same EA can increase considerably 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.

[4] The validity of (6) 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.

To fix ideas and to 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 homogenous across areas. This provides the simplest example of a (limited) failure of area homogeneity. More explicitly, we write

$$y_{ach} = \beta' X_{ach} + u_{ach} = \beta' X_{ach} + \eta_a + e_{ac} + \varepsilon_{ach}, \tag{7}$$

where $Cov(\eta_a, e_{ac}) = 0$, $Cov(\eta_a, \varepsilon_{ach}) = 0$, $Cov(e_{ac}, \varepsilon_{ach}) = 0$, $Cov(X_{ach}, u_{ach}) = 0$, $Cov(\varepsilon_{ach}, \varepsilon_{ach'}) = 0 \ \forall \ a, c, h, h' \neq h$, $Cov(e_{ac}, e_{ac'}) = 0 \ \forall \ a, c, c' \neq c$, $Cov(\eta_a, \eta_{a'}) = 0 \ \forall \ a, a' \neq a$. All error components are uncorrelated with each other and with the correlates. We assume that model (7) holds for every cluster c in region R, so that it also holds for all clusters within the small area. Model (7) 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 MCI fails because in a specific small area A:

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

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, $W_A = P(y \leq z \mid a = A)$, where z denotes the poverty line. The head count in A is equal to $P(y \leq z \mid a = A) = P(e_{ac} + \varepsilon_{ach} \leq z - \beta' X_{ach} - \eta_a)$, 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, but only of its expectation over the whole region R conditional on the correlates X, and depending on the actual value of η_a the two will generally be different at the level of the area A. Note also that even in the absence of area fixed effects, η_a, there can still be substantial MSE if the area does not include enough clusters to ensure that the impact of the cluster fixed effects e_{ac} on W_A averages to zero.

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 Error (MSE), and not on the sole variance of the point estimates. 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.

3.1 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 including a large number C of clusters labeled c, each of which includes a population of p households labeled h. For simplicity, and for this subsection only, we assume that both C and p 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 equi-correlated 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:

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

where ρ_a and ρ_c are respectively the intra-area (inter-cluster) and the intra-cluster correlation coefficients. In the specific case where the error term has a random effects structure as in (7), 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

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

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

$$\begin{aligned} M.S.E. &= E[(\hat{\mu}_y^a - \mu_y^a)^2 \mid a = A] = \eta_A^2 + \frac{\sigma_e^2}{C} + \frac{\sigma_\varepsilon^2}{Cp} \\ &= \eta_A^2 + \frac{\sigma_u^2}{Cp}[(\rho_c - \rho_a)p + (1 - \rho_c)] \\ &= \eta_A^2 + \frac{\sigma_u^2}{Cp}[1 + \rho_c(p-1)] - \frac{\sigma_\eta^2}{C}. \end{aligned} \tag{8}$$

The second term coincides with the variance of the estimator when the DGP in (7) does not include an area fixed effect, so that $\eta_A = \sigma_\eta = 0$. Both this and the third term converge to zero when the number of clusters in the small area becomes large, but the first term does not, and may lead to severe underestimation of the MSE in areas characterized by a large value of η_A.

Table 1 shows the underestimation of the root MSE for a given small area that would result assuming that area fixed effects are zero. We tabulate results for different parameter combinations,

keeping cluster size fixed at $p = 100$.[5] Each figure is the ratio between the (true) root MSE calculated as in (8) and the incorrect root MSE calculated assuming $\rho_a = \sigma_\eta = 0$, which is given by the second term in (8). For each combinations of ρ_c, ρ_a and C, we calculate ratios for two different values of the area fixed effect η, which are the 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^2_η and σ^2_ε are set, as is the distribution of η.

The results show that disregarding the bias component may 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 includes 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 fifty percent 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 percent 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 inter-cluster correlation becomes small relative to the intra-cluster correlation. Overall, the ratios in the table range from 1 to 19.9, both resulting from unlikely combinations that require a very high intra-cluster correlation equal to .20.

The MSE in (8) is calculated *conditional* on the area effect η. 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:

$$Var(\hat{\mu}_y) = \left(\frac{\sigma_u}{\sum_{c=1}^{C} p_c} \right)^2 \left\{ \sum_{c=1}^{C} p_c + \underbrace{\sum_{c=1}^{C} p_c(p_c - 1)\rho_c}_{\text{from intracluster corr.}} + \underbrace{\sum_{c=1}^{C} \sum_{c'=1, c' \neq c}^{C} p_c p_{c'} \rho_a}_{\text{from inter-cluster corr.}} \right\}. \tag{9}$$

The first term is the variance calculated assuming that observations are *i.i.d.*. The second and third term come respectively from the intracluster and inter-cluster correlation implied by model (7), because of the common geographical and socio-economic characteristics within the area that come from the failure of area homogeneity. This last term can be large. In the simple case where

[5] Results are much more sensitive to changes in C than to changes in p. Tabulations for different values of p are available upon request from the authors.

each cluster contains the same number of households, so that $p_c = p \ \forall c$, equation (9) simplifies to

$$\begin{aligned} Var(\mu_y - \hat{\mu}_y) &= \frac{\sigma_u^2}{Cp}\left[1 + (p-1)\rho_c + p(C-1)\rho_a\right] \\ &= Var_{SRS} + \frac{\sigma_u^2}{Cp}\left[(p-1)\rho_c + p(C-1)\rho_a\right] \\ &= Var_C + \frac{\sigma_u^2}{Cp}\left[p(C-1)\rho_a\right], \end{aligned} \tag{10}$$

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 zero as the number of clusters goes to infinity, the second term converges to $\rho_a\sigma_u^2$ which is not zero unless the intra-area (inter-cluster) correlation ρ_a is zero. 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 $C = 150$, $p = 100$, 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 fiftyfold.

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. Violation of this condition would lead to a further exaggeration of the correct MSE.

4 Estimation

In order to construct more realistic Monte Carlo experiments, we need a method for calculating poverty maps; here we propose a new estimator which we argue has practical advantages over the method usually used. As with the methods in ELL, we assume that both MP and CI hold. Given MCI, we can see from the modified moment condition (6) that the sampling process identifies the parameter of interest and we propose a non-parametric estimator based on the simple idea of replacing the modified moment condition (6) by its sample analog. To fix ideas, suppose that the object of interest is the head count poverty ratio in a small area A, calculated for a given fixed poverty line z. Under assumptions MP and CI, the head count can be estimated in two steps. In the fist step, $P(y \leq z \mid X)$ is estimated using survey data from region R. In the second step, the poverty count is calculated as $p_A^{-1}\sum_{h=1}^{p_A}\hat{P}(y_{Ach} \leq z \mid X_{Ach})$. We estimate the conditional probability non-parametrically, so as to avoid arbitrary assumptions about functional form. Note also that we directly model the conditional probability of being in poverty, instead of first estimating a model for the conditional distribution of y. The probability of being poor depends not only on the

conditional mean of y, but also on its conditional variance and higher moments, all of which need to be modeled or assigned values in an indirect approach; there is no need to model the complete conditional distribution when all we need is $P(y \leq z \mid X)$.

Using the notation laid out in Section 2, consider the more general framework where the object of interest is a welfare measure W_0. The estimate $\hat{W}_0$ is obtained as the solution to the following equation:

$$\frac{1}{p_A} \sum_{h \in H(A)} \hat{E}\left[s_h g(y_h; \hat{W}_0) \middle| X_h \right] = 0, \tag{11}$$

where p_A is the number of census households in the small area, and the expectation is estimated non-parametrically. Non-parametric estimation of a regression function in presence of many regressors can be cumbersome if carried out with standard methods—such as Nadaraya-Watson or locally linear regressions—that require the use of kernels.[6] However, the estimation can be greatly simplified using a sieve-based non-parametric estimator.[7] Sieve estimation of a conditional expectation is simple and fast, because it is implemented by straightforward least squares regression of the dependent variable on a series of "basis" functions of the regressors.

Formally, let $\{q_l(X), \quad l = 1, 2, ...\}$ denote a sequence of known basis functions. The basis functions are functions of X, such as powers or trigonometric functions of X, that can approximate any square-measurable function of X arbitrarily well. Also let n_R denote the number of observations sampled from region R in the household survey. Let $k(n_R)$ be an integer, with $k(n_R) \to \infty$ and $k(n_R)/n_R \to 0$ when $n_R \to \infty$. Finally, let

$$q^{k(n_R)}(X) = \left(q_1(X), ..., q_{k(n_R)}(X)\right)' \quad \text{and} \quad Q_R = \left(q^{k(n_R)}(X_{R1}), ..., q^{k(n_R)}(X_{Rn_R})\right)',$$

where the subscript R indicates observations from the household survey. So, $q^{k(n_R)}(X)$ is a $k(n_R) \times 1$ vector that includes the value of all the $k(n_R)$ basis functions evaluated at a specific point X, and Q_R is an $n_R \times k(n_R)$ matrix including the value of all the basis functions evaluated for each household in the survey at the corresponding observed values $X_{R1}, ..., X_{Rn_R}$. The first step of the sieve estimation can be written as the predicted values from an OLS regression,

$$\widehat{E}[s_h g(y_h; W) \mid X] = \sum_{h \in H_n(R)} s_h g(y_h; W) q^{k(n_R)}(X_{Rj}) \left(Q_R' Q_R\right)^{-1} q^{k(n_R)}(X). \tag{12}$$

In the second step, the projection for a given census observation with proxies X_{Ah} is calculated replacing X with X_{Ah} in (12), and $\hat{W}$ is obtained by grid search as the value of W that solves the equation. In practice, for many important welfare measures grid search is not necessary. For example, when W_0 is the head count ratio, the first step requires the estimation of the projection of a binary variable equal to one for poor households on a list of sieve basis functions, while in the

[6] Even so, notice that parametric rates of convergence for the parameter of interest can still be achieved, because W_0 is calculated as the integral of a conditional expectation, and not by the conditional expectation itself.

[7] For a simple introduction to sieve see Pagan and Ullah 1999, section 3.8.

second step $\hat{W}$ is simply the mean estimated probability of being poor for households in the small area, calculated using the coefficients estimated in the first stage.

Analogous estimation techniques have been developed in Chen et al. (2005) and Chen et al. (2007), who study the estimation of parameters defined by moment conditions with missing data in presence of auxiliary information. Such previous work has studied the properties of the estimators in the simplified context of *i.i.d.* observations, but consistency and asymptotic normality for the more general case of a clustered sample can be established as a straightforward extension. The calculation of the standard errors, which is described in detail in section 5, is simplified by the fact that the only sampling error derives from the first-stage estimation, because the predictors X used in the second step come from a census.

4.1 The ELL Estimator

The poverty maps constructed by the World Bank or with their assistance make use of an alternative estimation method proposed by Elbers et al. 2003 (ELL for brevity). Like our non-parametric estimator, ELL requires the validity of both MP and of the area homogeneity assumption CI. Unlike the estimator described in the previous section, however, ELL is a simulation-based estimator that uses strong parametric assumptions. 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 et al. (2003) and 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 $Cov(\eta_c, \varepsilon_{ch}) = Cov(X_c, \eta_c) = Cov(X_c, \varepsilon_{ch}) = Cov(\varepsilon_{ch}, \varepsilon_{c'h'}) = 0 \ \forall \ \ c, c', h, h' \neq h$. 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 function 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 $\hat{u}_{ch} - \widehat{\eta}_c$. The variance of the idiosyncratic error ε_{ch} is then estimated imposing the following parametric

form for heteroskedasticity:

$$\sigma^2(X) = \frac{Ae^{z'_{ch}\alpha} + B}{1 + e^{z'_{ch}\alpha}}, \tag{13}$$

where z_{ch} is a function of the correlates X, and A and B are parameters to be estimated. Using the estimates from (13) 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 of the parameters in (13). Second, each cluster in the census is assigned a cluster-specific error $\hat{\eta}^r_c$ drawn from the empirical distribution of all $\hat{\eta}$.[8] Third, each observation in the census is assigned a normalized idiosyncratic error e^{*r}_{ch} which is obtained either from a parametric distribution or from the empirical distribution of the errors. Fourth, heteroskedastic errors e^r_{ch} are calculated by using the parametric model in (13) evaluated at the simulated parameter values. Lastly, simulated values for $\ln y$ are generated as $\ln y^r = X'^r + \hat{\eta}^r_c + e^r_{ch}$, 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 $\widehat{Var}(\hat{W})$ (note the similarity with multiple imputation, Rubin 1987).

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 1). 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 non-parametric estimator requires no such assumptions. Both estimators, of course, require the absence of area heterogeneity, and this common ground is likely more important than their differences.

[8] This approach disregards the possible correlation among observations that belong to different clusters, and will therefore overstate the precision of the estimates. Such correlation would exist, for instance, if the true model includes an area fixed effects such as in (7). Demombynes et al. (2007) 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*. Based on the results of the empirical application in Elbers et al. (2002), they argue that when the intra-cluster correlation is small such conservative estimates of the standard errors will be only marginally different from those that assume no inter-cluster correlation. The arguments laid out in Section 3.1, as well as results from Monte Carlo experiments in the next section suggest instead that such upper bound can be enormously larger than the standard errors calculated under the assumption of no inter-cluster correlation.

5 Monte Carlo Experiments

We first consider a best-case scenario where the Data Generating Process (DGP) is characterized by a simplified version of model (7), where there is no small area fixed effect, the cluster fixed effects are *i.i.d.* and homoskedastic, and MP and CI hold. Specifically, for each cluster within a region the DGP is described as follows:

$$\begin{aligned} y_{ch} &= \beta_0 + \beta_1 x + u_{ch} = 20 + x_{ch} + e_c + \varepsilon_{ch} \\ x_{ch} &= 5 + z_{c,1} w_{ch} + z_{c,2}, \quad w \sim N(0,1), \quad z_{c,1}, z_{c,2} \sim U(0,1), \quad z_{c,1} \perp z_{c,2}, \\ e_c &\sim N(0, .01), \quad \varepsilon_{ch} \sim N(0, \sigma^2(x)) \\ \sigma^2(x) &= \frac{e^{\alpha_1 x + \alpha_2 x^2}}{1 + e^{\alpha_1 x + \alpha_2 x^2}}, \end{aligned} \tag{14}$$

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 (13). 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 (for instance 15-20,000 households), and the quantity to be estimated (for example 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 above 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 $p_A = 15,000$ 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 headcount 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 Monte Carlo 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, Root Mean Squared Error (RMSE) and confidence interval coverage rates ("coverage") 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 r^{th} Monte Carlo replication. The RMSE is estimated as the square root of

$R^{-1}\sum_{r=1}^{R}(\hat{W}_A^r - W_A)^2$, while 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 non-parametric and simulation-based ELL estimators described in Section 4. For a given auxiliary sample generated in the r^{th} replication, we calculate the nonparametric estimates as

$$\hat{W}_A(\hat{\gamma}) = \frac{1}{p_A}\sum_{h\in H(A)} \hat{E}\left[1(y_h < z)\left(1-\frac{y_h}{z}\right)^{\alpha}\Big| x_h\right] = \frac{1}{p_A}\sum_{h\in H(A)} \tilde{x}'\widehat{\gamma}, \tag{15}$$

where $\alpha = 0$ for the headcount ratio and $= 1$ for the poverty gap, $\tilde{x}$ are a sequence of basis functions, and $\widehat{\gamma}$ are the corresponding coefficients, estimated with Ordinary Least Squares using the artificial household survey data. As basis functions, we use a quadratic in x, $\sin(x)$, $\cos(x)$, $\sin(2x)$ and $\cos(2x)$. All Monte Carlo replications use the same artificial census population, which is therefore non-random, so that the only source of sampling error is the estimation of γ in the first stage. The variance of the parameters of interest can then be calculated as $G\widehat{Var}(\hat{\gamma})G'$, where $G \equiv (1/p_A)\sum_{h\in H(A)}\tilde{x}'$, and $\widehat{Var}(\hat{\gamma})$ is calculated taking into account the clustered nature of the sample. We note, however, that while the variance of the estimator only depends on the estimation error of the parameters γ, the Mean Squared Error will also be affected by the difference between W_A and $\hat{W}_A(\gamma)$. For instance, if W_A is a poverty head count ratio, a difference will generally exist, in a finite population, between the proportion of poor individuals and the mean value of the conditional probability of being poor. While in the following Monte Carlo experiments disregarding this bias components of the MSE is of little consequence, in Appendix B we describe circumstances where this is not true, and we propose an estimator of the MSE for such cases.

When adopting the ELL estimator, we estimate the heteroskedasticity parameters (α_1, α_2) using Non-Linear Least Squares, using the correct model (14) 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 4. 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 Monte Carlo replications. For all welfare measures, both estimators are essentially unbiased, and the RMSE is small relative to the true value being estimated. The RMSE is lower for the ELL estimator, presumably as a result of this estimator using more information than the non-parametric estimator, which does not use parametric assumptions (here assumed to be correct) about the form of the heteroskedasticity. Coverage rates are very close to the nominal ones (.95), and are actually above the nominal level for ELL, suggesting that this latter

estimator somewhat overestimates the true standard errors.[9]

Overall, when the parametric assumptions used by ELL are correct, both estimators perform well, although the non-parametric estimator is substantially simpler than ELL. The simulation-based estimator in ELL is also inconsistent if the model for heteroskedasticity is misspecified, while the NP estimator is robust to such mis-specifications.[10] 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.

5.1 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 (7). We still assume that there is no heterogeneity in the slope β that describes the conditional expectation of y given x. 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:

$$y_{ach} = 10 + 2x_{ach} + \eta_a + e_{ch} + \varepsilon_{ach}, \tag{16}$$

$$x \sim N(5,1) \quad \eta_a \sim N(0,\sigma_\eta^2) \quad e \sim N(0,\sigma_e^2) \quad \varepsilon \sim N(0,\sigma_\varepsilon^2). \tag{17}$$

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

$$P(y_{Ach} \leq z \mid A = a) = P(2x + e_{ch} + \varepsilon_{ach} \leq z - 10 - \eta_a) = \Phi\left(\frac{z - 10 - \eta_a - 15}{\sqrt{\sigma_e^2 + \sigma_\varepsilon^2 + 4}}\right),$$

where the last expression follows from the normality of the errors and the covariate x. As in the previous simulation, we consider fixed populations generated with the DGP described in (16) and (17). More specifically, we generate a synthetic census for three small areas with an area fixed effect η_a equal to the 50th, 75th and 99th percentile of the assumed distribution of η_a. When the area fixed effect becomes larger, we should expect the performance of both estimators to worsen, with coverage rates decreasing towards zero. Is should be noted that, given that the poverty measures considered here are non-linear functions of y, coverage rates will in general be incorrect even when we consider a small area with $\eta_a = 0$.

[9] This may be a result of the small number of observations per cluster (10) in the synthetic survey sample. ELL estimate the location effects as the intra-cluster means of the regression residuals (see Section 4.1). To illustrate, suppose that the coefficients β are known. Then the location effects would be estimated as $\eta_c + n_c^{-1}\sum_h \varepsilon_{ch}$, whose variance is $\sigma_\eta^2 + n_c^{-1}\sigma_{\varepsilon_{ch}}^2$, which overstates the true variance σ_η^2. Similarly, the variance of the idiosyncratic error ε_{ch} will be underestimated by $n_c^{-1}\sigma_{\varepsilon_{ch}}^2$. This implies that the overall variance of the residual will be estimated correctly, while the intra-cluster correlation will be overestimated. The net effect would be an overestimate of the variance of the welfare measure estimates. This bias disappears when n_c is large, but most household surveys include only 10-20 units selected from the same cluster.

[10] Monte Carlo simulations that study the performance of ELL when a misspecified model for heteroskedasticity is imposed are available upon request from the authors.

We assume that the region from which the auxiliary data set is drawn is composed of 25 small areas, and that each small area is composed of 15,000 households split among 150 equally sized EAs. 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}, $i = 1, ..., 25$, $p_i = 0.01 + (.98/24) \times (i-1)$, 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 zero (the median and mean of the distribution), and when $i = 25$ it is equal to the 90th percentile.

Monte Carlo results for two alternative models based on 200 replications are displayed in Table 3. Each ELL estimation is obtained with 150 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 1000 households is generated from the DGP in (16) and (17). We draw four EAs from each one of the 25 small areas, and then we draw 10 "households" from each EA. For each DGP, the objects of interest are head count ratios calculated for poverty lines corresponding to the 10th and 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 form simpler than described in Section 4.1. The first step is unchanged, and consists of the estimation of model (16) using Ordinary Least Squares, followed by the 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 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.

When we use the non-parametric estimator, we estimate $\hat{P}(y < z \mid x)$ using a sieve logit estimator, where we use a series of basis functions $\tilde{x}$ as regressors (see e.g. Hirano, Imbens, and Ridder, 2003 or Chen, Hong, and Tarozzi, 2007). In this Monte Carlo, the first stage dependent variable is binary, and a binary dependent variable model such as logit has the advantage of producing predicted values which are, by construction, included in the unit interval. In the second step, the estimate of the head count ratio for an area A is calculated as the mean value of the predicted probability of being poor in the area. Because the census population is kept fixed and is therefore non-random, the only source of sampling error is the estimation of the logit parameters. Hence the variance of the estimate can be calculated using the delta method (see, e.g. Wooldridge 2002a, pp. 44-45) as follows:

$$\widehat{G}\widehat{Var}(\hat{\beta})\widehat{G}', \tag{18}$$

where

$$\widehat{G} \equiv \frac{1}{p_A} \sum_{h \in H(A)} F(\tilde{x}_h'\hat{\beta})(1 - F(\tilde{x}_h'\hat{\beta}))\tilde{x}_h', \qquad F(\tilde{x}_h'\hat{\beta}) = \frac{e^{\tilde{x}_h'\hat{\beta}}}{(1 + e^{\tilde{x}_h'\hat{\beta}})^2},$$

and where $\widehat{Var}(\hat{\beta})$ is the estimated covariance matrix of the first-stage coefficients, calculated taking into account the clustered survey design.

The top half of Table 3 shows a first set of results, where the DGP implies moderately large intracluster correlation (.11) and inter-cluster correlation (.06). Both the non-parametric estimator (columns 1-4) and ELL (columns 5-8) perform well in predicting the poverty counts when the area fixed effect is zero (panel 1A). Bias is always very small, and in this case ELL performs better than the non-parametric estimator both in terms of MSE and in terms of coverage. In some cases, the standard errors as calculated with ELL appear in fact to be conservative, so that coverage rates are close to one.[11] The performance of both estimators worsens dramatically 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 percent of the mean value of the "expenditure" variable y), the coverage rate for both estimators always remains well below .25, and is actually very close to zero in several cases (panel 1B). The results in panel 1C show that when the area fixed effect is .818 (the 90th percentile of the distribution of η) the coverage rate decreases to zero in all cases. Consistently with the results in Section 3, the decline in coverage is caused by the increase in bias associated with the presence of the area fixed effect. While in panel 1A virtually all MSE derives from the standard error of the estimator (because the bias is close to zero), in panels 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 using a modified version of ELL where—in each of the 150 simulations required to complete one Monte Carlo replication—the *same* cluster fixed effect is assigned to all households within the same area (see Elbers et al. (2002)). This modification should lead to 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. As a consequence, within each Monte Carlo replication, the standard error as calculated via simulations will be larger than the true one.[12] In fact, this modified methodology leads to standard errors which are approximately ten times as large as those estimated in column 6. The increase in the standard errors now leads to confidence intervals which always include the true value, so that coverage rates are equal to one in all cases, even when the area fixed effect is relatively large (panel 1C). However, the confidence intervals are now so wide to become barely informative. For instance, "standard" ELL produced for $P_0(20)$ a confidence interval of width 0.045 ($0.01137 \times 1.96 \times 2$), while "conservative" ELL produces intervals of width .51 ($.13073 \times 1.96 \times 2$).

In the bottom half of Table 3, we show that coverage rates may be far from the nominal 95% even in cases where intra-cluster correlation is very small. The DGP in Model 2 implies a small

[11] See footnote 9.

[12] This is also the reason why the standard errors reported in column 11 of Table 3 are *larger* than the MSE. The estimated MSE is the sum of the *true* bias and variance as calculated across Monte Carlo simulations, while in column 11 we report the mean value of the standard errors as calculated by the "conservative" ELL procedure.

intra-cluster correlation (.0178), but also implies that most of it is due to the presence of the area fixed effect, so that the inter-cluster correlation is 0.0153. As a result, coverage rates decline rapidly for both ELL and the non-parametric estimator, and when the area fixed effect is moderately large (panel 1C) coverage approaches or equals zero. This result is consistent with the figures in Table 1, where we have shown that when the ratio between inter and intra-cluster correlation is large, standard errors which do not take inter-cluster correlation into account will severely underestimate the true MSE, leading to misleading inference. Overall, disregarding the presence of area fixed effects may lead to severely misleading inference even when the intra-cluster 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.[13]

Note also that we have assumed that unobserved heterogeneity is only present in the intercept, while we have assumed a DGP with a constant slope for all observations that belong to the same region. In real empirical applications, it is likely that inference will be further 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.[14]

6 An Empirical Evaluation Using Census Data from Mexico

The Monte Carlo experiments in Section 5 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. 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 and Sobek 1997). Like most census micro-data, the 2000 Mexican Census includes many predictors of income/expenditure, such as housing characteristics, household composition, asset ownership, occupation and education of each household member. Unlike most census data sets, however, this census also includes a measure of individual income during the previous 30 days. This allows us to carry out an experiment which 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 income (y) and of a set of predictors (x). We use this sample to estimate the parameters of a model for the probability of being poor (that is, of income being below a fixed poverty line z) conditional on

[13] Coverage rates approach the nominal ones when the inter-cluster correlation is relatively small. Monte Carlo results are available upon request.

[14] As for the case of heterogeneity in the intercepts, coverage rates will diverge more from the nominal ones when the variability of the coefficients is large relative to their mean value. Results from Monte Carlo results analogous to those described in this section are available upon request.

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.[15] 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 likely severely 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 aggregating villages *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 largely removed by construction. In our case, areas 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*. The details of the validation experiment are described in Section 6.1, while the reader only interested in the results can refer directly to Section 6.2.

It should be noted that 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. Indeed a non-negligible fraction of households report zero income over the previous 30 days. However, a comparison between census 2000 and the 2002 Mexican Family Life Survey (MFLS) does not show major discrepancies. In rural Oaxaca, median monthly per capita income was 80.8 Pesos according to MFLS, and 94.8 according to census 2000. In the state of Veracruz, the MFLS median was 233.3 Pesos, while the census estimate was 238.[16]

6.1 Details of the Validation Exercise

The complete IPUMS micro-data extract for Mexico 2000 includes more than ten million observations, so to keep the validation exercise manageable we limit our analysis to the rural section of three of the largest Mexican states, that is, 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. To illustrate, the map in Figure 1 shows the subdivision of the state of

[15] As described in Section 6.1, we assume that the true values are identical to the estimates obtained from the census extract.

[16] We do not report on a comparison for Chiapas because this state was not separately identified in MFLS 2002.

Chiapas into *municipios* according to the 2005 Geo-statistical Census of Mexico.[17] 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 MFLS ranged from 47 (Distrito Federal) to 469 (in Michoacán).[18] 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 *municipios* is very heterogeneous but relatively large. The median household population size of the rural sector of a *municipio* is 2829 in Chiapas, 443 in Oaxaca and 2150 in Veracruz (see Table 4). Hence our choice of using *municipios* as small areas.

Because we wish to work with a census, while IPUMS only includes a 10.6% extract of the complete micro-data, 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 (non-random) 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. Such strategy is suggested in ELL to reduce the extent of intra-cluster 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 that the researcher has access only to a (pseudo) household survey defined here as ten observations from each of 50 *municipios* selected at random without replacement. By construction, this sampling design leads to different probability of selection for different households, so that estimation is done after construction of sampling weights. We classify a household as poor if total monthly income per head y is below a threshold z equal to 200 Pesos.[19] Because the census actually includes income for all households, the true value of the headcount ratio for each *municipio* can be calculated as in (3) with $\alpha = 0$, that is, as the proportion of individuals who live in households with per capita income below z. These true headcount ratios can then be compared with the corresponding estimates obtained using synthetic survey samples and making use of a (possibly incorrect) conditional independence assumption.

We evaluate the coverage of 95% nominal confidence intervals via 250 Monte Carlo simulations. We complete independent simulations for the three states of Chiapas, Oaxaca and Veracruz.

[17] See http://www.cuentame.inegi.gob.mx.. The subdivisions in 2000 and 2005 were essentially identical.

[18] For these calculation we have classified households as rural when they live in communities with population below 2,500.

[19] The USD Mexican Peso PPP exchange rate in 2000 was 6.79, so that 200 Pesos corresponds to approximately one PPP dollar per person per day (Heston et al. 2006).

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 5 provides a list of the predictors in the first stage.[20]

Once the artificial sample has been selected, we estimate $\hat{P}(y < z \mid x)$ using a logit model, as described in Section 3.1, and then we estimate the poverty head count ratios for each *municipio* as the mean predicted probability of being poor. The latter is calculated using the parameters estimated in the first step together with census information on the predictors for all households in the *municipio*. The confidence intervals are calculated as $\hat{W} \pm 1.96 \times \widehat{RMSE}$, where $\hat{W}$ is the point estimate and $\widehat{RMSE}$ is an estimate of the root-MSE calculated as described in Appendix B. Finally, we record if the true value of the headcount ratio in each *municipio* lies within the interval boundaries.

6.2 Results of the Validation Exercise

The predictors listed in Table 5 have moderately good explanatory power is each of the three states. Using all complete observations from the census, the pseudo-R^2 is .1868 in Chiapas, .1694 in Oaxaca, and .1636 for Veracruz.[21] For each state, we display in Figure 1 histograms of coverage rates. Each observation shows, for a given *municipio*, the fraction of Monte Carlo replications for which the true value of the poverty headcount ratio lies within a nominal 95% confidence interval. Because assumption MP holds by construction, deviation of the coverage rates from the nominal ones likely indicates failure of the homogeneity assumption CI. As a reminder, our estimation procedure does not rely on distributional assumptions, nor does it rely on parametric assumption about the form of heteroskedasticity.

The histograms in Figure 2 show that while coverage rates for most areas are not far from the nominal 95%, there is a large fraction of areas where 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 CI ignores. The fraction of *municipios* where coverage remains below .75 is .36 in Chiapas, .54 in Oaxaca and .49 in Veracruz. In all the three states coverage rates are below 50% in approximately 10 percent 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 .61), .40 in

[20] 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.

[21] The full estimation results are available upon request.

Oaxaca (minimum .22 and maximum 1.89) 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, Monte Carlo 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 and Table 6). 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 which 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 extents of heterogeneity will be present elsewhere, although the plausibility of spatial heterogeneity in intercepts or in 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*.

7 Conclusions

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, the 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. Such methodology is deemed able to allow for the estimation of welfare estimates for areas of less 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 the necessary conditions for usefully matching survey and census data are unlikely to hold in practice. Second, for situations when such conditions hold, we describe a non-parametric estimator that is faster, simpler, and more robust than the estimator proposed in Elbers et al. 2003. Finally, we argue that both methodologies 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 heterogeneity across small areas in the conditional distribution of expenditure or income given a series of predictors. The presence of area heterogeneity is apparent in an empirical experiment carried out with data from the 2000 Mexican census.

Overall, we believe that efforts to calculate welfare estimates for small areas merging survey and census data are certainly worthwhile, but we also believe that the current literature has not

emphasized enough the limitations of the current methodologies and the very strong assumptions that they require in order to allow for meaningful inference. Such limitations should be stressed, and the precision of the estimates should be judged accordingly. Policy makers that make use of small area statistics to allocate funds and improve targeting of welfare programs should be aware that such maps may be subject to much more uncertainty and error than previously thought.

References

Chen, S. and M. Ravallion (2004). How have the world's poorest fared since the early 1980s? *The World Bank Research Observer 19*(2), 141–169.

Chen, X., H. Hong, and E. Tamer (2005). Measurement error models with auxiliary data. *Review of Economic Studies 72*(2), 343366.

Chen, X., H. Hong, and A. Tarozzi (2007). Semiparametric efficiency in GMM models with auxiliary data. *Annals of Statistics Forthcoming*.

Citro, C. F. and G. Kalton (Eds.) (2000). *Small-area income and poverty estimates. Priorities for 2000 and beyond*. Washington, DC.: National Academy Press.

Cochrane, W. G. (1977). *Sampling techniques*. New York: John Wiley.

Deaton, A. and M. Grosh (2000). Consumption. In M. Grosh and P. Glewwe (Eds.), *Designing household survey questionnaires for developing countries: lessons from 15 years of the Living Standards Measurement Study*, Volume 1, Chapter 5, pp. 91–133. Oxford University Press for the World Bank.

Demombynes, G., C. Elbers, J. O. Lanjouw, and P. Lanjouw (2007). How good a map? Putting small area estimation to the test. World Bank Policy Research Working Paper 4155.

Dorfman, R. (1979). A formula for the Gini coefficient. *Review of Economics and Statistics 61*(1), 146–149.

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

Elbers, C., J. O. Lanjouw, and P. Lanjouw (2002). Micro-level estimation of welfare. Policy Research Working Paper 2911, The World Bank, Washington, DC.

Grosh, M. and J. N. K. Rao (1994). Small area estimation: an appraisal. *Statistical Science 9*(1), 55–93.

Heckman, J., R. LaLonde, and J. Smith (1999). The economics and econometrics of active labor market programs. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics, Vol. 3A*. Amsterdam, The Netherlands: Elsevier Science.

Heston, A., R. Summers, and B. Aten (2006). Penn World Table version 6.2. Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania. http://pwt.econ.upenn.edu/php_site/pwt_index.php.

Hirano, K., G. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

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

Lee, L. and J. Sepanski (1995). Estimation of linear and nonlinear errors-in-variables models using validation data. *Journal of The American Statistical Association 90*(429), 130–140.

Little, R. and D. Rubin (2002). *Statistical Analysis with Missing Data* (Second ed.). New York: John Wiley & Sons.

National Research Council (1980). *Panel on small-area estimates of population and income. Estimating population and income of small areas*. Washington, DC.: National Academy Press.

Pagan, A. and A. Ullah (1999). *Nonparametric Econometrics*. Cambridge, UK: Cambridge University Press.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology 66*, 688–701.

Rubin, D. (1976). Inference and missing data. *Biometrika 63*, 581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

Ruggles, S. and M. Sobek (1997). Integrated public use microdata series: Version 2.0. Historical Census Projects, University of Minnesota. http://www.ipums.umn.edu.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*(2), 461–464.

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

Wooldridge, J. (2002b). Inverse Probability Weighted M-estimators for sample selection, attrition and stratification. *Portuguese Economic Journal 1*, 117–139.

APPENDIX A - PROOFS

Proof of equation (6):

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

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

Proof of equation (9):

$$Var(\hat{\mu}_y) = Var(\mu_y - \hat{\mu}_y) = \left(\frac{1}{\sum_{c=1}^{C} p_c}\right)^2 Var\left(\sum_{c=1}^{C}\sum_{h=1}^{p_c} u_{ch}\right)$$

$$= \left(\frac{1}{\sum_{c=1}^{C} p_c}\right)^2 \left\{\sum_{c=1}^{C}\sum_{h=1}^{p_c}\sigma^2 + \sum_{c=1}^{C}\sum_{h=1}^{p_c}\left(\sum_{c'=1,c'\neq c}^{C}\sum_{h'=1}^{p_{c'}}\rho_a\sigma^2\right) + \sum_{c=1}^{C}\sum_{h=1}^{p_c}\left(\sum_{h'=1,h'\neq h}^{p_c}\rho_c\sigma^2\right)\right\}$$

$$= \left(\frac{\sigma}{\sum_{c=1}^{C} p_c}\right)^2 \left\{\sum_{c=1}^{C} p_c + \sum_{c=1}^{C}\sum_{h=1}^{p_c}\left(\sum_{c'=1,c'\neq c}^{C} p_{c'}\rho_a\right) + \sum_{c=1}^{C} p_c(p_c-1)\rho_c\right\}$$

$$= \left(\frac{\sigma}{\sum_{c=1}^{C} p_c}\right)^2 \left\{\sum_{c=1}^{C} p_c + \sum_{c=1}^{C}\sum_{c'=1,c'\neq c}^{C} p_c p_{c'}\rho_a + \sum_{c=1}^{C} p_c(p_c-1)\rho_c\right\}.$$

APPENDIX B - Bias of the Nonparametric Estimator

The estimated standard errors of the nonparametric estimator proposed in this paper are calculated treating the estimation of the first-stage parameters as the sole source of sampling error. For instance, if the parameter of interest is a headcount ratio, the estimator is

$$\hat{W}_A = \frac{1}{p_A} \sum_{h \in H(A)} \hat{E}\left[1(y_h < z) | \, x_h\right] = \frac{1}{p_A} \sum_{h \in H(A)} F(\tilde{x}'\widehat{\gamma}), \tag{19}$$

where $\tilde{x}$ are a sequence of basis functions and $\widehat{\gamma}$ are coefficients, estimated using survey data with OLS (if $F(.)$ is the identity function) or logit (if $F(.)$ is the logistic cdf). The variance of the parameters of interest is then calculated as $\hat{G}\widehat{Var}(\hat{\gamma})\hat{G}'$, where $\hat{G}$ is the estimated gradient. However, this estimator ignores the fact that, in reality, the true head count ratio would be identical to the mean value of the conditional probability of being poor only if the small area included an infinite number of households. However, in general we will have that

$$W_A = \frac{1}{p_A} \sum_{h \in H(A)} 1(y_h \leq z) \neq \frac{1}{p_A} \sum_{h \in H(A)} P(y_h \leq z \mid x_h). \tag{20}$$

If the bias that derives from the above difference is disregarded, coverage rates may be different from the nominal ones. By definition, the MSE of the estimator is equal to $\left[bias\left(\hat{W}_A\right)\right]^2 + var\left(\hat{W}_A\right)$. As described in the paper, the variance is estimated as

$$\left[\frac{1}{p_A} \sum_{h \in H(A)} \frac{\partial F\left(\tilde{x}_h'\hat{\gamma}\right)}{\partial \gamma}\right] \widehat{var}\left(\hat{\gamma}\right) \left[\frac{1}{p_A} \sum_{h \in H(A)} \frac{\partial F\left(\tilde{x}_h'\hat{\gamma}\right)}{\partial \gamma}\right]. \tag{21}$$

By definition, the bias for a *given area* A is:

$$\begin{aligned} bias\left(\hat{H}\right) &= E\left[\frac{1}{p_A} \sum_{h \in H(A)} F\left(\tilde{x}_h'\hat{\gamma}\right)\right] - \frac{1}{p_A} \sum_{h \in H(A)} 1\left(y_h \leq z\right) \\ &= \frac{1}{p_A} \sum_{h \in H(A)} \left[E\left(F\left(\tilde{x}_h'\hat{\gamma}\right)\right) - 1\left(y_h \leq z\right)\right]. \end{aligned}$$

This quantity is unknown, and it depends on the specific small area being considered. Here we describe an estimation procedure and Monte Carlo experiments which evaluates its performance. First, we approximate $E\left[F\left(\tilde{x}_h'\hat{\gamma}\right)\right]$ with the true value $P\left(y_h \leq z | \, x_h; \gamma\right)$. Letting $p_h \equiv$

$P\left(y_h \leq z | \, x_h; \gamma\right)$, we have then

$$\begin{aligned} bias^2\left(\hat{H}\right) &\approx \left[\frac{1}{p_A}\sum_{h\in H(A)}\left(p_h - 1\left(y_h \leq z\right)\right)\right]^2 \\ &= \frac{1}{p_A^2}\sum_{h\in H(A)}\sum_{h'\in H(A)}\left[p_h - 1\left(y_h \leq z\right)\right]\left[p_{h'} - 1\left(y_{h'} \leq z\right)\right] \\ &= \frac{1}{p_A^2}\sum_{h\in H(A)}\left[p_h - 1\left(y_h \leq z\right)\right]^2 \\ &\quad + \frac{1}{p_A^2}\sum_{h\in H(A)}\sum_{h'\in H(A), h'\neq h}\left[p_h - 1\left(y_h \leq z\right)\right]\left[p_{h'} - 1\left(y_{h'} \leq z\right)\right]. \end{aligned} \tag{22}$$

Both the two terms on the right-hand side of (22) include the value of y_h for all observations in the area and are therefore unknown. We estimate $bias^2\left(\hat{H}\right)$ by replacing the elements in the summations with sample estimates of their expected value. So, letting

$$\begin{aligned} r_1 &= E\left[\left(p_h - 1\left(y_h \leq z\right)\right)^2\right] \\ r_2 &= E\left[\left(p_h - 1\left(y_h \leq z\right)\right)\left(p_{h'} - 1\left(y_{h'} \leq z\right)\right)\right], \end{aligned}$$

the estimated squared bias for a given small area is then:

$$\widehat{bias}^2\left(\hat{H}\right) = \frac{\hat{r}_1}{p_A} + \frac{p_A - 1}{p_A}\hat{r}_2. \tag{23}$$

Note that while the first term in (23) goes to zero when the size of the area increases, the second term does not.

We perform a series of Monte Carlo experiments to explore the performance of the nonparametric estimator with or without inclusion of the bias estimation in (23). We assume that expenditure can be modeled as:

$$\begin{aligned} y_{ah} &= 20 + \beta x_{ah} + \eta_a + \varepsilon_{ah}, \\ x_{ah} &\sim N\left(5, 1\right), \quad \varepsilon_{ah} \sim N(0,1), \quad \eta_a \sim N\left(0, \sigma_\eta^2\right) \end{aligned}$$

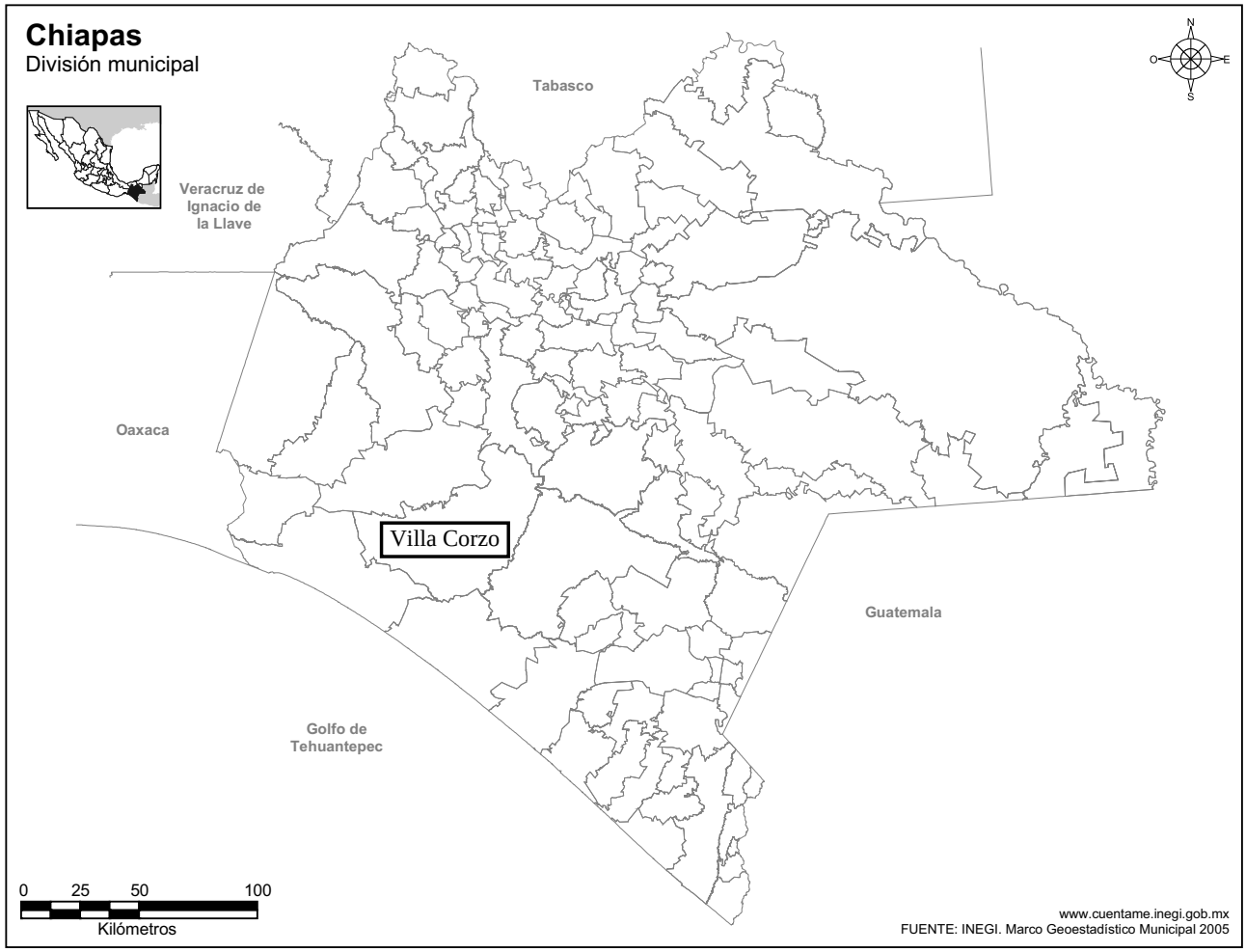
where x, η and ε are assumed independent. The DGP mimics the framework relevant for the pseudo-validation exercise with Mexican data in Section 6. We experiment with different models where we vary both the size of the small area and the magnitude of the R^2 and of the intra-cluster correlation coefficient $\rho = \sigma_\eta^2/(\sigma_\eta^2 + \sigma_\varepsilon^2)$. Given that σ_ε^2 is kept equal to one, 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

$$R^2 = 1 - \frac{var(\eta_a + \varepsilon_{ah})}{var(y_{ah})} \Rightarrow \beta = \sqrt{\frac{R^2\left(1 + \sigma_\eta^2\right)}{\left(1 - R^2\right)}}.$$

We experiment with a range of R^2 and ρ in a range consistent with values usually found in 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 $p_A \in \{100, 500, 5000, 15000\}$. Each coverage rate is calculated over 1,000 Monte Carlo 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 1000 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 25^{th} percentile of the distribution of y when the area fixed effect η_a is zero, 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 6. It is apparent that if the bias is not taken into account (columns 1, 3, 5 and 7) the nonparametric estimator systematically underestimate the true prediction error, so that coverage rates are almost always below .95, often by a lot. 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 intra-cluster 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 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-.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 $\hat{r}_2$ in (23) often 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.

Figure 1: Map of *Municipios* in the State of Chiapas (Mexico)



Source: INEGI, Mexico. The map illustrates the 119 *municipios* that form the state of Chiapas according to the 2005 geo-statistical census of Mexico. The 2000 Census of Mexico listed 6519 households in the *municipio* of Villa Corzo highlighted in the map.

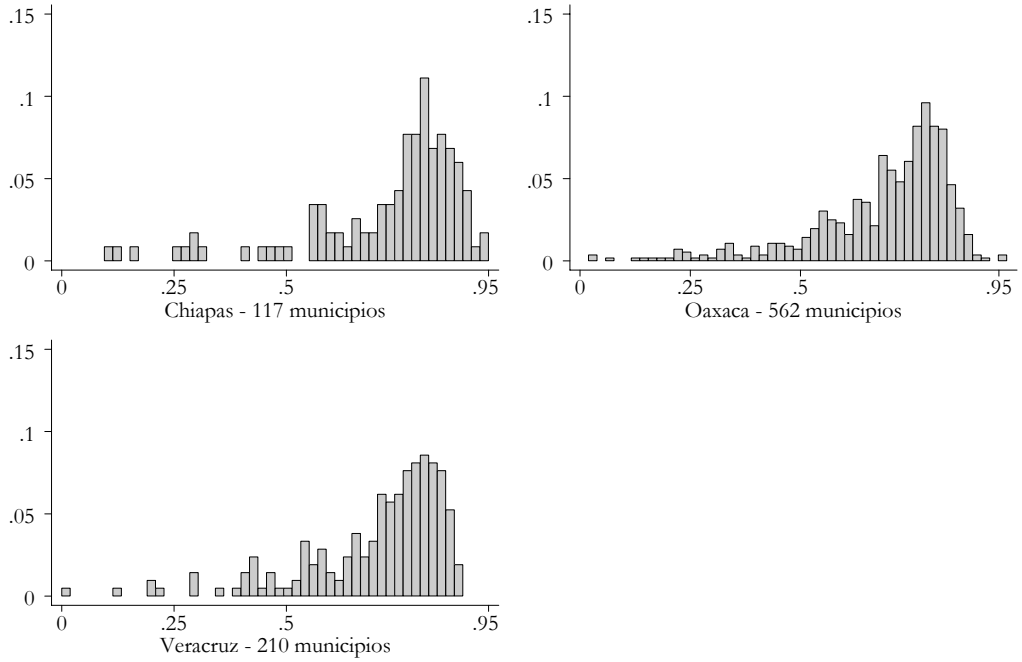


Figure 2: Distribution of Coverage Rates by state, Poverty Head Counts

Source: authors' calculations from IPUMS Mexico 2000 Census extract. An observation indicates, for a given *municipio*, the fraction of Monte Carlo replications for which the true value of the poverty headcount ratio lies within a nominal 95% confidence interval.

Table 1: Effect of Inter-cluster 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 intra-cluster correlation (ρ_c), Inter-cluster 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 includes 100 households. Given a probability distribution function $f(\eta)$ for the area fixed effect η, the value $\tau_{\eta,p}$ is the p-th 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.

Table 2: Monte Carlo Simulations - No Inter-cluster Correlation

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Nonparametric Estimator			ELL		
	True Value	Bias	RMSE	Coverage	Bias	RMSE	Coverage
$P_0(24)$	0.0979	0.0013	0.0101	.944	0.0015	0.0079	.988
$P_0(25)$	0.3323	-0.0032	0.0165	.924	-0.0031	0.0129	.972
$P_0(26)$	0.6732	-0.0068	0.0154	.912	-0.0056	0.0128	.968
$P_1(24)$	0.0023	0.0000	0.0003	.920	0.0000	0.0003	.972
$P_1(25)$	0.0103	0.0001	0.0007	.944	0.0001	0.0006	.988
$P_1(26)$	0.0292	0.0000	0.0010	.944	0.0000	0.0010	.988

Notes: 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 1000 households selected from 100 equally-sized clusters.

Table 3: Monte Carlo Simulations - Consequences of Inter-cluster Correlation

		Non-parametric				**ELL**				**ELL, Conservative s.e.**			
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
		Bias	s.e.	RMSE	Covrg	Bias	s.e.	RMSE	Covrg	Bias	s.e.*	RMSE	Covrg
	True Value												
		Model 1 - $\sigma_\eta = .5, \quad \sigma_e = .5, \quad \sigma_\varepsilon = 2. \quad R^2 = .471, \quad \rho_c = .111, \quad \rho_a = .056$											
		(1A) - $\eta_a = \tau_{\eta,.50} = 0$											
$P_0(16.3)$	0.096	0.00572	0.00970	0.01124	0.875	0.00531	0.00709	0.00885	0.955	0.00635	0.06095	0.01065	1
$P_0(18.1)$	0.246	0.00518	0.01367	0.01458	0.900	0.00588	0.01038	0.01190	0.980	0.00738	0.10615	0.01503	1
		(1B) - $\eta_a = \tau_{\eta,.75} = .329$											
$P_0(16.3)$	0.078	0.02384	0.00970	0.02573	0.205	0.02343	0.00709	0.02448	0.135	0.02447	0.06095	0.02592	1
$P_0(18.1)$	0.212	0.03989	0.01367	0.04215	0.085	0.04059	0.01038	0.04189	0.085	0.04208	0.10615	0.04407	1
		(1C) - $\eta_a = \tau_{\eta,.90} = .818$											
$P_0(16.3)$	0.056	0.04580	0.00970	0.04681	0.000	0.04539	0.00709	0.04594	0.000	0.04643	0.06095	0.04721	1
$P_0(18.1)$	0.166	0.08576	0.01367	0.08684	0.000	0.08646	0.01038	0.08708	0.000	0.08796	0.10615	0.08893	1
		Model 2 - $\sigma_\eta = .25, \quad \sigma_e = .10, \quad \sigma_\varepsilon = 2. \quad R^2 = .496, \quad \rho_c = .0178, \quad \rho_a = .0153$											
		(2A) - $\eta_a = \tau_{\eta,.50} = 0$											
$P_0(16.3)$	0.098	0.00353	0.0094	0.01002	0.905	0.00299	0.00623	0.0069	0.955	0.00371	0.04435	0.00810	1
$P_0(18.1)$	0.249	0.00376	0.01273	0.01324	0.910	0.00391	0.00856	0.00939	0.975	0.00487	0.07810	0.01157	1
		(2B) - $\eta_a = \tau_{\eta,.75} = .165$											
$P_0(16.3)$	0.089	0.01323	0.00940	0.01622	0.645	0.01270	0.00623	0.01414	0.620	0.01342	0.04435	0.01523	1
$P_0(18.1)$	0.231	0.02183	0.01273	0.02525	0.560	0.02198	0.00856	0.02358	0.530	0.02294	0.07810	0.02522	1
		(2C) - $\eta_a = \tau_{\eta,.90} = .409$											
$P_0(16.3)$	0.075	0.02630	0.00940	0.02792	0.115	0.02576	0.00623	0.02650	0.015	0.02649	0.04435	0.02744	1
$P_0(18.1)$	0.205	0.04724	0.01273	0.04892	0.025	0.04739	0.00856	0.04816	0.000	0.04835	0.07810	0.04948	1

Notes: 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 1000, 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 p^{th} 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. See Section 5.1 for simulation details.

Table 4: Mexico 2000 Pseudo-census: Summary Statistics

	Chiapas	Oaxaca	Veracruz
Extract size (no. households)	56554	118328	94480
Pseudo-census hhs. population size	385816	393256	606899
Pseudo-census individual population size	2052071	1897684	2834599
no. of *municipios*	118	570	210
Mean no. of hhs. in a pseudo-census *municipio*	3270	700	2890
Median no. of hhs. in a pseudo-census *municipio*	2700	413	1991
Fraction of households reporting zero income	16.4	20.2	12.5
Fraction of individuals with missing income	1.97	1.88	1.88

Source: authors' calculations from Mexico 2000 Census IPUMS extract (rural only). See Section 6.1 for details about the construction of the pseudo-census.

Table 5: 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/stone
Primary roof material is masonry/concrete/tile
Speaks only indigenous language
Speaks both indigenous language and Spanish
Head is working
Head works in Agriculture/Fishery/Forestry/Mining
household members ages 0-12 (and its squared)
household members older than 65 (and its squared)
male members ages 13-65 (and its squared)
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/stone
Primary roof material is masonry/concrete/tile
Speaks only indigenous language
Head works in Agriculture/Fishery/Forestry/Mining

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

Table 6: Coverage of Nonparametric Estimator

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	$\rho = 0$		$\rho = .02$		$\rho = .05$		$\rho = .10$	
	s.e.	Bias	s.e.	Bias	s.e.	Bias	s.e.	Bias
	10 hhs 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 = 5000								
$R^2 = .20$	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 = 15000								
$R^2 = .20$	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 hhs from 1000 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 = 15000								
$R^2 = .20$	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 1000 Monte Carlo replications. Confidence intervals are calculated either taking into account only the standard error of the estimator ("s.e.") or including also an estimate of the bias squared ("bias"). 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 25^{th} percentile of the distribution of y when the area fixed effect η is zero. 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 indicated the size of the small area population, which is generated in each replication as a random draw from the DGP. * refer to the number of areas from which a given number of synthetic households is drawn.