

Can Census Data Alone Signal Heterogeneity in the Estimation of Poverty Maps?

Alessandro Tarozzi
Duke University

April 2008*

Abstract

Methodologies now commonly used for the construction of poverty mapping assume a substantial degree of homogeneity within geographical areas in the relationship between income and its predictors. However, local labor and rental markets and other local environmental differences are likely to generate heterogeneity in such relationships, at least to some extent. The purpose of this paper is to argue that useful even if only indirect and suggestive evidence on the extent of area heterogeneity is readily available in virtually any census. Such indirect evidence is provided by non-monetary indicators—such as literacy, asset ownership or access to sanitation—which are routinely included in censuses. These indicators can be used to perform validation exercises to gauge the extent of heterogeneity in their distribution conditional on predictors analogous to those commonly used in poverty mapping. We argue that the same factors which are likely to generate area heterogeneity in poverty mapping are also likely to generate heterogeneity in such kind of validation exercises. We construct a very simple model to illustrate this point formally. Finally, we evaluate empirically the argument using data from the 2000 Mexican census, which also included measures of income. In our empirical illustration, the performance of imputation methodologies to construct maps of indicators typically feasible with census data alone is informative about how effectively such methodologies are able to produce poverty maps with the correct precision. **JEL: I32, C31**

Key words: Census, Heterogeneity, Poverty Maps

*I am grateful to Angus Deaton and Barbara Rossi for useful comments and suggestions and to Maria Genoni for excellent research assistance. I am also grateful to IPUMS for granting access to the 2000 Mexican Census extract. I am solely responsible for all errors and omissions. Address: Dept of Economics, Duke University, Social Sciences Building, PO Box 90097, Durham, NC 27708, taroz@econ.duke.edu.

1 Introduction

Estimates of poverty for virtually all countries in the world are typically calculated using income or expenditure data recorded in household surveys. The sample size and design of these surveys are often such that relatively precise poverty estimates can also be obtained for sub-national administrative units such as states or provinces. However, the estimation of welfare indicators becomes usually impossible or very imprecise if one is interested in poverty for small geographical areas such as towns or districts. For such small areas, a nationally representative household survey will typically provide a handful of observations, or none at all. On the other hand, certain socio-economic indicators such as literacy or access to sanitation can usually be determined even for very small areas during census years, when the whole population is surveyed, at least in principle. However, in most countries, censuses do not collect income or expenditure information, so that small area poverty estimates are typically not available even in census years.

To address this shortcoming, [Elbers et al. \(2003\)](#) have developed an imputation procedure that merges information from a census and a household survey with the purpose of estimating precise—that is, low variance—welfare measures at the local level. Variations of such procedure have been applied 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 Intuitively, suppose that a survey contains a “large” number of observations from a given *region* for which both income/expenditure (“income” hereafter) and some predictors such as asset ownership or literacy are measured. Suppose also that the same predictors are similarly measured in a census. Then it may be possible to recover precise estimates of welfare measures at the local level by merging information on the conditional distribution of income given the predictors estimated from the survey with census information on the distribution of the predictors.

The methodology introduced by [Elbers et al. \(2003\)](#) first uses survey data to estimate a projection of (log) income on a set of predictors and to estimate the properties of the residuals. Simulations are then used to estimate local welfare measures making use of the parameter estimates from the first step and census information on the predictors at the local level. [Tarozzi and Deaton \(2008\)](#) describe a simpler alternative projection-based estimator and they show that its performance is similar to that of the methodology in [Elbers et al. \(2003\)](#) under a variety of conditions. When, for instance, the researcher is interested in estimating small area-specific poverty headcounts, the projection estimator proceeds in two simple steps. First, the probability of being poor conditional on the chosen predictors is estimated using survey data collected from a relatively large geographical region. Then the poverty head count for a given area within

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.

the region is calculated as the mean probability of being poor for individuals in the small area. The second step is completed by using the coefficients estimated in the first step together with the predictors recorded in the census for all units in the small area.

Clearly, in order for either of these methodologies to allow correct statistical inference, one must assume a substantial degree of homogeneity within the region in the relationship between income and its predictors. Whether such homogeneity assumption holds is ultimately an empirical matter. [Tarozzi and Deaton \(2008\)](#) argue that the presence of local labor and rental markets and other local environmental differences are likely to generate heterogeneity which will violate this assumption, at least to some extent. Using data from the 2000 Mexican census, they show that disregarding heterogeneity leads to the estimation of confidence intervals for poverty headcount rates that do not have correct coverage. In other words, confidence intervals constructed for a poverty count H_0 using a “nominal” 95% coverage, that is, calculated as $\hat{H} \pm 1.96 \times \widehat{M.S.E.}(\hat{H})$, included in fact the true value H_0 less than 95% of the times. The validation exercise was possible because the Mexican census, unlike most censuses in developing countries, includes measures of income. It was then possible to calculate “true” poverty counts at the local level, which were compared to estimates obtained through small area estimation techniques using data from synthetic “household surveys” generated from the census. A similar validation methodology is used in [Elbers et al. \(2008\)](#) who, using data from the Brazilian state of Minas Gerais, find that the coverage of confidence intervals constructed for small areas have coverage close to nominal.²

Of course, the extent of area heterogeneity is likely to differ in different countries and regions. As a consequence, the incorrect coverage rates found in Mexico do not *imply* that incorrect coverage should be expected elsewhere. Similarly, correct coverage found somewhere does not imply that correct coverage should be expected everywhere. Unfortunately, direct assessments of the reliability of poverty maps using validations are not generally possible, because they require the availability of income measures in the census whose absence is the very reason why small area estimation techniques are necessary. The purpose of this paper is to argue that useful, if only indirect, evidence on the extent of area heterogeneity is readily available in virtually any census. Such indirect evidence is provided by non-monetary indicators—such as literacy, asset ownership or work status—which are routinely included in censuses. Validation exercises can be completed using such indicators to gauge the extent of area heterogeneity in their conditional distribution given predictors analogous to those typically used in poverty mapping. It is important to

²Note, however, that [Elbers et al. \(2008\)](#) base their results on only 41 Monte Carlo replications, completed by pooling together samples drawn according to three separate sampling schemes. Most of the samples are drawn using a peculiar stratification scheme which is not clearly taken into account in the calculation of the standard errors. Also, the authors do not document how the predictors are chosen and include, in their model to predict income, mean values of income itself. Income means at the local level would clearly not be available in a typical poverty mapping exercise.

stress that such evidence will in any case only be indirect, and *cannot and should not* be used as a formal test on the accuracy of poverty mapping. To illustrate, the existence of area homogeneity of the conditional probability of being poor given, say, literacy and employment status *does not imply and is not implied by* homogeneity across areas of the conditional probability of being literate given employment status.

We argue, however, that differences across areas in relative prices, relative returns and other location-specific characteristics are likely to generate heterogeneity in the distribution of any wealth-related indicator conditional on other such indicators. To illustrate this point formally, we construct a simple model where individuals allocate an exogenously determined income between three different goods, with individual-specific preferences and area heterogeneity in relative prices. The solution of the model shows that the same sources of heterogeneity appear in both the conditional expectation of (log) income given expenditure in two of the consumed goods and in the conditional expectation of (log) expenditure on one good given consumption in the second good. Obviously, the model does not pretend to be realistic. However, it shows that because most variables used in poverty mapping (e.g. income, asset ownership, education, dwelling characteristics etc.) are the result of choices made by households within the constraints of local prices and returns and other local unobservable conditions, the resulting heterogeneity, when present, is likely to characterize most or all conditional distributions that involve such variables. A series of Monte Carlo simulations where the data generating process is consistent with the simple model shows that ignoring heterogeneity leads to confidence intervals with coverage below the nominal level.

We evaluate the empirical relevance of our argument using a census microdata sample from the 2000 Mexican census. Unlike in most countries, the Mexican census also includes measures of income. [Tarozzi and Deaton \(2008\)](#) use these data to replicate a poverty mapping exercise, assuming that one has access only to a synthetic “survey” sampled from the census, and then comparing the estimated maps with the values reconstructed from the full census extract. They find that coverage rates are below the nominal level for a large number of small areas, which suggests the presence of a sizeable degree of area heterogeneity. Here, we employ analogous simulation techniques to evaluate the performance of mapping when the outcomes of interest are based on variables which are routinely measured in most censuses. Namely, we estimate “literacy” and “sanitation” maps. In the former case the statistic of interest is the fraction of the population in an area who live in households with a literate head. In the latter case, we estimate the fraction of individuals with access to a toilet. According to the argument put forth in this paper, incorrect coverage rates in the estimation of poverty maps would also be reflected in similarly incorrect coverage in the estimation of maps of other indicators associated to poverty. We find that this is indeed the case. In the context of this empirical application, the performance of non-income maps constructed using only information usually available in a census *is* informative about the performance of poverty mapping.

The rest of the paper is organized as follows. The next section briefly summarizes the statistical problem and describes the estimator used in [Tarozi and Deaton \(2008\)](#). Section 3 describes a simple model to illustrate why heterogeneity in some structural relations among household-level economic variables is likely to signal the presence of heterogeneity in different structural relations. Section 4 describes the pseudo-experiment with data from the 2000 Mexican Census, and Section 5 concludes.

2 Statistical Background and Estimation

The purpose of this section is to describe the statistical framework of the construction of “maps” of welfare measures. Because our main purpose here is not to describe techniques for small area estimation, we only provide a simplified description for the case where the welfare measure is the fraction of units (e.g. households or individuals) in a given area with a given characteristic. The most common example is perhaps the poverty head count, that is, the fraction of the population with income per head below a given threshold (the poverty line). More generally, the same framework could be applied to the estimation of other statistics. For instance, a policy maker could be interested in evaluating the extent of illiteracy, unemployment or malnutrition across different areas (see for instance [Fuji 2007](#)). A formal and more general framework which also encompasses other poverty or inequality measures is described in [Tarozi and Deaton \(2008\)](#), from which this section borrows. See also [Elbers et al. \(2003\)](#) for an alternative methodology.

Suppose that the object of interest is the head count ratio H for a “small area” A , where $A \subset R$ denotes a small area such as a town or district included in a larger “region” R such as 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 about 100 households. Let H_i denote a binary variable equal to one if individual i has a given characteristic, such as being poor, unemployed or illiterate. Let also $\Theta(A)$ denote the set of individuals who live in area A . The parameter of interest is then

$$H_A = \frac{1}{N_A} \sum_{i \in \Theta(A)} H_i, \tag{1}$$

where N_A is the number of elements in $\Theta(A)$.³ When data on H_i are recorded in a survey, the head count H_A can be trivially estimated by its sample analogue $\bar{H}_A = n_A^{-1} \sum_{i \in \Theta_n(A)} H_i$, where $\Theta_n(A)$ indicates the set of individuals in the survey sampled from area A , and n_A indicates their number. However, a survey may include few or no observations at all on H_i if the area A is sufficiently small. We assume instead that the survey sample from the broader region R includes a relatively large number of individuals for which

³For simplicity, we also abstract from the difference between individual and household specific welfare measures.

(H_i, \mathbf{x}_i) are observed, where \mathbf{x}_i indicates a vector of household or community characteristics correlated with H_i . We also assume that \mathbf{x}_i is measured in the census as well, so that it is observed for all $i \in \Theta(A)$. The set of predictors \mathbf{x}_i can also include community characteristics or location-specific means of household-level variables recorded only in the census, as long as the survey includes detailed geographical identifiers which allow to merge such census information into the survey (Elbers et al. 2003). When instead census and survey provide separate measurements of the same variables, care should be taken to make sure that the information has been collected consistently across the two data sources, because it is well-known that data collection can be heavily influenced by the methodology adopted to elicit information (Deaton and Grosh 2000).

To fix ideas, it is now useful to adopt a superpopulation approach, assuming that the small area A includes an infinite number of individuals, so that using the law of iterated expectations we can write

$$H_A = E(H_i | i \in \Theta(A)) = E[P(H_i = 1 | \mathbf{x}_i, i \in \Theta(A)) | i \in \Theta(A)]. \quad (2)$$

Even if the survey sample contains no observations on H_i , the head count can be estimated using auxiliary information from the census if one can assume $P(H_i = 1 | \mathbf{x}_i, i \in \Theta(R)) = P(H_i = 1 | \mathbf{x}_i, i \in \Theta(A))$. Such “stability” assumption, labeled *area homogeneity* in Tarozzi and Deaton (2008), is closely related to the conditional independence assumptions which has been used extensively in statistics and econometrics to address missing data problems due to non-response, attrition, measurement error or unobserved counterfactuals in program evaluation.⁴ As emphasized in Tarozzi and Deaton (2008), area homogeneity is a demanding assumption in the estimation of small area statistics. Suppose for instance that H_A is a poverty head count and \mathbf{x} includes schooling or occupation variables. Heterogeneity in the conditional probability of being poor given \mathbf{x} is likely to arise due to differences across areas in the local rates of return. The presence in \mathbf{x} of measures of asset ownership will likely capture some of the variation in the rates of return, but assets are subject to similar concerns because their relative prices and rates of return may vary across areas. In general, area homogeneity will fail when the relevant conditional probability is a function of unobserved factors such as relative prices or returns, tastes or other location-specific characteristics which vary across different areas. Concerns are also likely to become more severe when the census and the survey used for poverty mapping are completed in different years. This is a common circumstance, because while a census is usually completed only once every ten years, household surveys are often completed at shorter intervals. In such cases, the construction of poverty maps has to rely on information merged from two data sets collected possibly years apart from each other. This is problematic, especially for developing countries whose economies are often growing and changing rapidly.

⁴For extensive references see, for instance, Carroll et al. (1995), Heckman et al. (1999), Little and Rubin (2002), Rubin (1996), Todd (2007), Chen et al. (2008).

Keeping all these *caveats* in mind, area homogeneity, coupled with the availability of predictors \mathbf{x}_i in both census and survey, provide a basis for a two-step estimation of the welfare measure in the small area A based on the sample analogue of (2). First, the parameters of the binary dependent variable model $P(H_i = 1 \mid \mathbf{x}_i; \theta)$ are estimated using survey data from region R . Second, the estimated coefficients $\hat{\theta}$ are used to form imputed values of the conditional probabilities for all census observations from the small area, and the head count for the small area is finally calculated as

$$\hat{H}_A = \frac{1}{N_A} \sum_{i \in \Theta(A)} P(H_i = 1 \mid \mathbf{x}_i; \hat{\theta}). \quad (3)$$

In this paper, as in [Tarozzi and Deaton \(2008\)](#), we model the conditional probability $P(\cdot)$ using logit.⁵ The standard errors of \hat{H}_A can be estimated in a straightforward way using the delta method, recalling that because the census population is kept fixed and is therefore non-random, the only source of sampling error derives from parameter estimation. Hence:

$$\widehat{Var}(\hat{H}_A) = \hat{\mathbf{G}} \widehat{Var}(\hat{\theta}) \hat{\mathbf{G}}' \quad (4)$$

where

$$\hat{\mathbf{G}} \equiv \frac{1}{N_A} \sum_{i \in \Theta(A)} F(\tilde{\mathbf{x}}'_i \hat{\theta}) (1 - F(\tilde{\mathbf{x}}'_i \hat{\theta})) \tilde{\mathbf{x}}'_i,$$

where $\tilde{\mathbf{x}}_i$ is the vector of functions of \mathbf{x}_i chosen as regressors and $F(\cdot)$ denotes the cumulative distribution function used in the estimation of the conditional probability. When such probability is estimated using a logit model, $F(\tilde{\mathbf{x}}'_i \hat{\theta}) = e^{\tilde{\mathbf{x}}'_i \hat{\theta}} (1 + e^{\tilde{\mathbf{x}}'_i \hat{\theta}})^{-1}$.

In the common circumstance that the survey data used for the estimation of $\hat{\theta}$ are collected using a two-stage clustered design, the standard errors of the parameters should be adjusted taking intra-cluster correlation into account.⁶ It is important to note, however, that the expression in (4) may substantially underestimate the mean squared error (MSE) of the estimator if the number of observations in the small area A is not large. In fact, while \hat{H}_A is a consistent estimator for the area-specific mean value of $P(H_i = 1 \mid \mathbf{x}_i; \theta)$, the true object of interest is (1). In a superpopulation approach the two quantities are identical, but in real empirical applications their difference (that is, the *bias* of the proposed estimator) may be different enough to make inference based on (4) incorrect. Intuitively, because poverty mapping is

⁵Note that this approach can be made robust to parametric misspecification by estimating the conditional probability using a non-parametric estimator such as sieve-logit (see [Hirano et al. 2003](#) or [Chen et al. 2008](#)). However, non-parametric estimation is complicated by the necessity of choosing smoothing parameters and the degree of flexibility may be limited by sample size.

⁶More specifically, if θ is estimated using logit, robust standard errors can be easily produced using built-in commands in most statistical packages. For instance, in Stata this can be done using the `svylogit` command, or the `cluster` option in the standard `logit` command.

essentially an exercise in forecasting, confidence intervals for \hat{H}_A should be calculated taking into account not just the variance of the estimator, but rather its MSE, which also takes into account the (squared) bias. Letting $p_i \equiv P(H_i = 1 \mid \mathbf{x}_i; \theta)$, [Tarozzi and Deaton \(2008\)](#) show that the squared bias $b^2(\hat{H}_A)$ can be approximated by the following expression

$$\hat{b}^2(\hat{H}_A) = \frac{\hat{E}[(p_i - H_i)^2]}{N_A} + \frac{N_A - 1}{N_A} \hat{E}[(p_i - H_i)(p_{i'} - H_{i'})], \quad (5)$$

where both expectations are estimated using their respective sample analogues. Let $\Lambda_n(R)$ indicate the set of all areas a in region R included in the sample and let $\Theta_n(a)$ indicate the set of the n_a individuals sampled from area a (and define $\Theta_n(R)$ analogously), then

$$\begin{aligned} \hat{E}[(p_i - H_i)^2] &= \frac{1}{\sum_{a \in \Lambda_n(R)} n_a} \sum_{i \in \Theta_n(R)} (\hat{p}_i - H_i)^2 \\ \hat{E}[(p_i - H_i)(p_{i'} - H_{i'})] &= \frac{1}{\sum_{a \in \Lambda_n(R)} n_a (n_a - 1)} \sum_{a \in \Lambda_n(R)} \sum_{i \in \Theta_n(a)} \sum_{i' \in \Theta_n(a), i' \neq i} (\hat{p}_i - H_i)(\hat{p}_{i'} - H_{i'}). \end{aligned}$$

To summarize, a confidence interval for \hat{H}_A with nominal coverage $(1 - \tau)$ will be constructed as

$$\hat{H}_A \pm \Phi^{-1}(1 - \tau/2) \times \left[\widehat{Var}(\hat{H}_A) + \hat{b}^2(\hat{H}_A) \right], \quad (6)$$

where $\Phi^{-1}(\cdot)$ is the inverse of the cumulative distribution function of a standard normal. In a series of Monte Carlo experiments, [Tarozzi and Deaton \(2008\)](#) show that confidence intervals constructed in this way have approximately correct coverage with a variety of data generating processes, even when the number of observations within the area A is as small as 100 individuals.

3 A Formal Illustration

In this section we describe a very simple model to illustrate formally the idea that differences across areas in relative prices will generate heterogeneity in the distribution of wealth-related indicators conditional on other such indicators. We assume that individuals allocate an exogenously determined income between three different goods, with heterogeneity in preferences across individuals and heterogeneity in prices across areas. Within the boundaries of such oversimplified model, we show that heterogeneity in the distribution of expenditure in one of the goods conditional on expenditure in a second good is reflected also in heterogeneity of the distribution of income given expenditure in both goods. Both conditional distributions are in fact functions of relative prices, which we assume to be heterogeneous across different “areas”. More generally, our purpose is to illustrate that if factors such as relative price variation across areas lead to area heterogeneity, such heterogeneity is likely to be reflected not only in area heterogeneity

in the probability of being poor given predictors, but also in heterogeneity in the distribution of each predictor given the remaining ones.

Suppose that each area a is populated by individuals i with different nominal income Y_i , and suppose also that the budget is allocated among three goods, whose relative prices differ across different areas. We assume that preferences are described by a simplified version of the cost function that leads to an Almost Ideal Demand System (AIDS, [Deaton and Muellbauer 1980](#)). Specifically, we assume that the cost function C for individual i in area a is

$$\ln C_i(u, \mathbf{p}^a) = \alpha_i(\mathbf{p}^a) + ub_i(\mathbf{p}^a), \quad (7)$$

where $\mathbf{p}^a \equiv [p_1^a \ p_2^a \ p_3^a]$ is the vector of area-specific prices, u is a reference utility level, and

$$\begin{aligned} \alpha_i(\mathbf{p}^a) &= \sum_{k=1}^3 \alpha_{ik} \ln p_k^a \\ b_i(\mathbf{p}^a) &= \prod_{k=1}^3 (p_k^a)^{\beta_{ik}}, \end{aligned}$$

where homogeneity of degree zero requires that $\sum_{k=1}^3 \alpha_{ik} = 1$ and $\sum_{k=1}^3 \beta_{ik} = 0$.⁷ The existence of heterogeneity in relative prices across areas implicitly requires that markets are not perfectly integrated, and that each area can be seen as a separate economy. In some cases this maybe be a very strong assumption, but in reality such differences in relative prices are often observed, especially in developing countries where market integration is limited. The solution of the cost minimization problem leads to the following demand functions for goods 1 and 2

$$x_{i1} = \frac{Y_i}{p_1^a} \alpha_{i1} + \beta_{i1} \frac{Y_i}{p_1^a} \ln \left(\frac{Y_i}{P_i^a} \right) \quad (8)$$

$$x_{i2} = \frac{Y_i}{p_2^a} \alpha_{i2} + \beta_{i2} \frac{Y_i}{p_2^a} \ln \left(\frac{Y_i}{P_i^a} \right) \quad (9)$$

where $x_{ij}, j = 1, 2$ indicates the quantity of good j consumed by individual i , and $\ln P_i^a = \alpha_{i1} \ln p_1^a + \alpha_{i2} \ln p_2^a + \alpha_{i3} \ln p_3^a$ is an individual-specific price index. Suppose that an individual is counted as poor if $Y_i < z_y$, that is, if his/her income remains below a given poverty line z_y . In this model, relative prices will affect the probability of being poor conditional on the quantities consumed x_{i1} and x_{i2} . Similarly, differences in relative prices across different areas will generate differences across areas in the shape of the conditional probability $P(x_{i1} < z_{x_1} \mid x_1)$, where z_{x_1} can be interpreted as a ‘‘poverty line’’ in terms of good 1. Hence, heterogeneity in relative prices will likely lead to incorrect inference if a unique model for the conditional probability is imposed for all areas. The next subsection uses a Monte Carlo simulation to further illustrate this point.

⁷Using the notation in [Deaton and Muellbauer \(1980\)](#), this is a special case of the AIDS cost function where we have assumed $\alpha_{i0} = \beta_{i0} = \gamma_{i,km} = 0$, where the parameters γ allow for cross-substitution.

3.1 A Monte Carlo Experiment

Assume that the economy in a given region is described by individuals with the preferences described in (7). We assume that income Y_i is exogenous and generated from a normal distribution, with $Y_i \sim N(1000, 40000)$. We generate individual-specific preference parameters assuming that both α_{1i} and α_{2i} are drawn independently from a normal distribution with mean $1/4$ and standard deviation 0.05 , while β_{1i} and β_{2i} are drawn independently from a normal with mean 0.05 and standard deviation 0.005 . The remaining parameters α_{3i} and β_{3i} are calculated from the price homogeneity condition. This DGP ensures that income and demand are always positive in our simulations. We also assume that p_1^a , p_2^a and p_3^a are *area-specific* and drawn from a tri-variate normal distribution with means equal to 10 , standard deviations equal to σ_p and correlations equal to 0.3 . We always choose a variance small enough to ensure that the probability of drawing negative prices is essentially zero. Once income, prices and preferences are determined, consumption levels of goods 1 and 2 are calculated using (8) and (9). The case where $\sigma_p = 0$ corresponds to a situation where there is no area heterogeneity in relative prices.

Each Monte Carlo experiment proceeds as follows. First we generate data as described above for the synthetic “census” of a region composed of 50 small areas with 10,000 individuals each. We assume that the object of interest is either the fraction of individuals in each area with income below a poverty line $z_y = 1000$, or the fraction in each area with consumption of x_1 below a threshold $z_{x_1} = 50$. The census is then kept fixed and treated as the actual population. In each of 200 Monte Carlo replications we draw a sample of 1,000 individuals from the synthetic census, selecting 50 individuals from 20 small areas. Both individuals and areas are selected using simple random sampling without replacement. Once the synthetic survey has been generated, the conditional probabilities $H_y \equiv P(Y_i < z_y \mid x_{i1}, x_{i2})$ and $H_{x_1} \equiv P(x_{i1} < z_{x_1} \mid x_{i2})$ are estimated using a logit model, and point estimates and MSE of head counts are estimated as described in (3) and (6). After each Monte Carlo simulation, we determine whether the value of each area-specific mean (calculated from the synthetic census) lies within the estimated 95% confidence interval. Finally, we calculate for each area the coverage rates, that is, the fraction of the confidence intervals which includes the value calculated from the census.

In Figures 1 and 2 we show histograms of the estimated distribution of coverage rates, where each of the 50 observations is represented by a different area. If the estimator is asymptotically normal and the MSEs are estimated correctly, coverage for all areas should be approximately equal to the nominal size of the confidence intervals, that is, .95. In such case, the histograms would simply show a spike at or close to .95.

By construction, the DGP described above produce poverty head count rates H_y approximately equal to .50 in all areas. When there is no heterogeneity in prices, that is, if $\sigma_p = 0$, the estimated root-MSEs

(RMSE) for H_y are relatively small and range—depending on replication and area—from .008 to .088, with a mean of .025. When relative prices are common across areas, the head counts H_{x_1} are also by construction approximately constant and equal to 0.35. The estimated root-MSEs range from .010 to .099, with a mean equal to 0.033. As expected, coverage rates for both parameters of interest are approximately correct for all 50 areas, as shown in Figure 1.

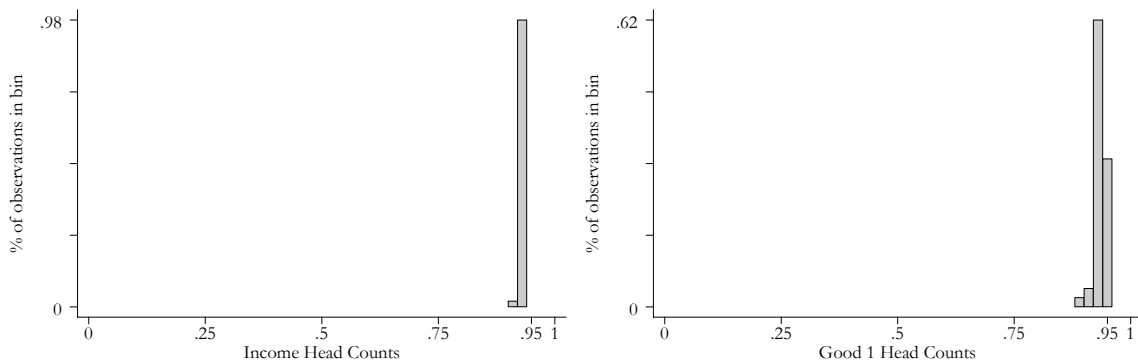


Figure 1: Histograms of coverage rates, based on 200 Monte Carlo replications. No area heterogeneity in relative prices ($\sigma_p = 0$). Each observation represents a different “small area” of 10,000 synthetic units.

The results change considerably when heterogeneity in prices is introduced. In panels (A) and (B) of Figure 2 we show histograms for the distribution of coverage rates for H_y and H_{x_1} respectively, when the standard deviation of prices σ_p is equal to 0.1. This modification now produces variation in the area-specific values of H_{x_1} , because demand for good 1 now changes depending on its relative price. In this simulation, the “good 1 head counts” range from .31 to .40. Note that the increased value of σ_p still leads to fairly homogeneous relative prices. For instance, the ratio p_1^a/p_2^a in the 50 synthetic areas ranges from a minimum of .96 to a maximum of 1.02. Even so, coverage rates for both parameters now worsen, even if in almost all cases they remain above .75.

Panels (C) and (D) of Figure 2 display the corresponding histograms for coverage rates when the standard deviation of prices is increased to 0.5. This produces more heterogeneous relative prices, now ranging from .82 to 1.12. This in turn leads to area-specific good 1 head counts ranging from .15 to .54. Coverage rates worsen further, to the point of becoming smaller than .75 for a large fraction of the small areas. This happens despite the fact that the estimated MSEs are now more than twice as large as for the case where $\sigma_p = 0$. The mean value of RMSE is now .059 for H_y and .082 for H_{x_1} .

To summarize, at least in the context of this very simple model, poor coverage in “good 1 poverty maps” signals poor coverage for income poverty maps, because heterogeneity in the underlying conditional probabilities is due to a common cause, that is, heterogeneity in relative prices. The model is obviously a naive representation of a true economy, but it should still provide a useful exemplification of the main

argument put forth in this paper: that is, that unobserved heterogeneity in local characteristics such as prices is likely to be reflected both in heterogeneity in the distribution of income given some of its correlates and in heterogeneity in the distribution of such correlates conditional on the others. In the next section, we use data from the 2000 Census of Mexico to study if such an argument is consistent with evidence from a real empirical setting.

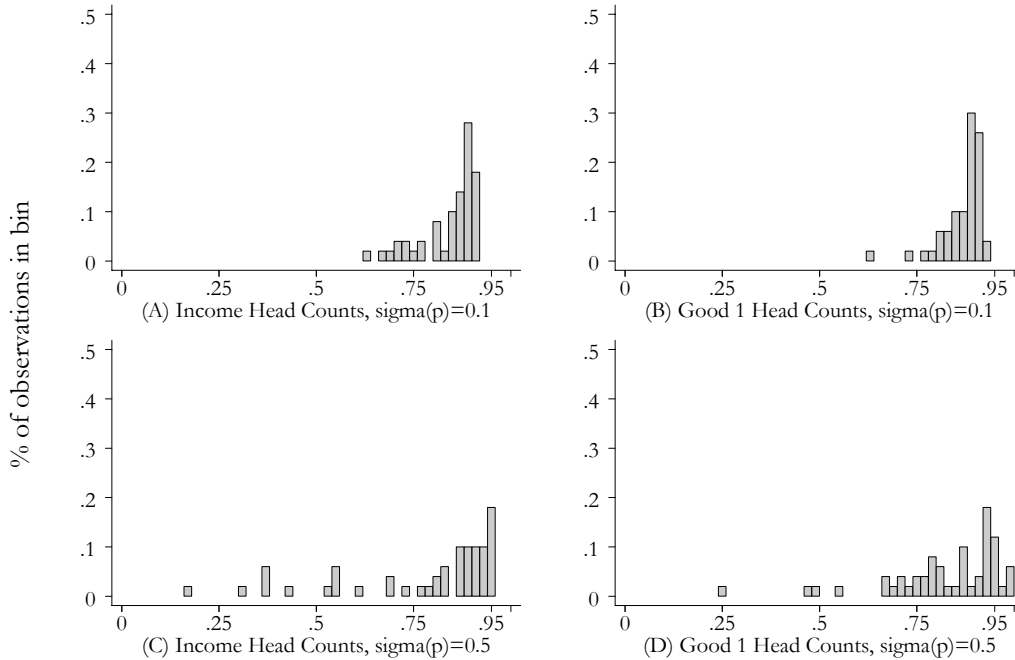


Figure 2: Histograms of coverage rates, based on 200 Monte Carlo replications. The magnitude of heterogeneity in relative prices is indicated by σ_p , denoted “sigma(p)” in the graphs. Each observation represents a different “small area” of 10,000 synthetic units.

4 An Illustration with Data from the 2000 Mexican Census

This section evaluates the merit of the arguments put forth in the previous section by using data from the 10.6% random extract of the 2000 Mexican Census from the Integrated Public Use Micro Sample (IPUMS).⁸ As customary for most census micro-data, the 2000 Mexican Census includes a long list of correlates of income. Among the others, such covariates include housing characteristics, household composition, asset ownership, occupation and schooling indicators of each household member. Unlike most census data sets, however, Mexico 2000 also includes a measure of individual income during the previous 30 days. This allows [Tarozzi and Deaton \(2008\)](#) to compare the poverty rates calculated with the *actual* income data

⁸See [Ruggles and Sobek \(1997\)](#), [Minnesota Population Center \(2007\)](#).

with those estimated simulating a poverty mapping procedure carried out assuming that only a random sample of observations is available from the census. The results in [Tarozzi and Deaton \(2008\)](#) suggest the existence of substantial heterogeneity in the areas considered. In fact, while poverty mapping produces overall informative results, the estimated MSEs appear to be biased downwards for a large number of small areas. Here we show that an analogous under-coverage of confidence intervals can be observed if poverty mapping methodologies are applied to the calculation of maps for alternative economic indicators, where both the indicators and the predictors are variables typically measured in census data. The indicators we choose are the fraction of individuals living in households where the head can read and write in at least one language (“literacy head counts”) and the fraction of individuals with access to a toilet (“sanitation head counts”). As in [Tarozzi and Deaton \(2008\)](#), we use data from three of the largest Mexican states, namely Chiapas, Oaxaca and Veracruz. We treat each state as a separate region (R), and we adopt smaller administrative units called *municipios* as small areas (A). Because the census extract includes only a handful of observations from some *municipios*, we only use observations from *municipios* with at least 500 households. Table 1 shows that this brings the number of small areas to 108 in Chiapas, 240 in Oaxaca and 182 in Veracruz.

Table 1: Mexico 2000 Pseudo-census: Summary Statistics

	Chiapas	Oaxaca	Veracruz
Pseudo-census hhs. population size	395078	318701	612942
Pseudo-census individual population size	2035661	1528763	2793627
no. of <i>municipios</i>	108	240	182
Mean no. of hhs. in a pseudo-census <i>municipio</i>	3658	1328	3367
Poverty Head Count Ratio (Pov. line 200 Pesos day/person)	.66	.65	.44
Literacy Head Count	.68	.70	.72
Sanitation Head Count	.71	.74	.78

Source: IPUMS Mexico Census 2000. All figures are calculated including only *municipios* with at least 500 households in the complete census. These *municipios* account for 99 percent of the population in Chiapas (which includes a total of 118 *municipios*), 81 percent in Oaxaca (562 *municipios*) and 99 percent in Veracruz (210). The “literacy” and the “sanitation” head counts are defined as the fraction of the population living in households with, respectively, a literate household head and access to a toilet.

All the three states have high poverty rates. The mean area-specific head count is .66 in Chiapas, .65 in Oaxaca and .44 in Veracruz. The mean fraction of literate household heads is similar across states and close to 70 percent, while the mean proportion of individuals with access to a toilet ranges from 71 percent in Chiapas to 78 percent in Veracruz. Figure 3 shows that there is substantial degree of variation in the heterogeneity of these statistics across states. In all states, the *municipio*-level distribution of poverty head counts covers most of the unit interval and is relatively uniform, with more evidence of left-skew in Chiapas and Oaxaca, the two poorest states. The range of the sanitation head counts distributions is similarly wide,

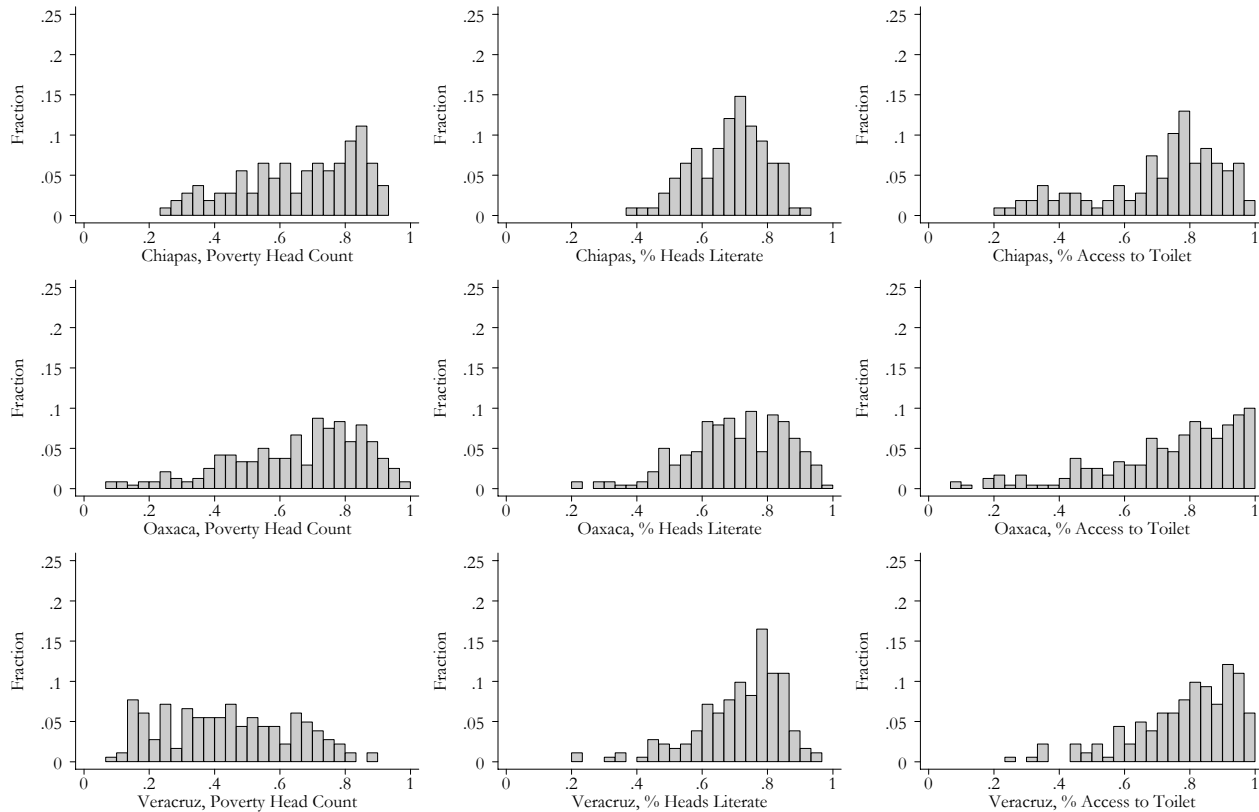


Figure 3: Distributions of *municipio*-specific statistics, by state.

and all distributions are left-skewed, including in Veracruz. The distributions of the literacy head counts are left-skewed too, but appear to be somewhat more concentrated around their mean.

Table 2 contains the list of variable included as predictors of poverty status. Individuals are considered to be poor when income per head is below 200 Pesos per month.⁹ Note that the list of predictors also includes location means. As in Elbers et al. 2003, the inclusion of location means typically reduces the correlation of the residuals in the first-stage regression, which in turn reduces the bias of the estimator described in Section 2. In many circumstances, such location means can be included in the mapping procedure even if they refer to variables which are *only* recorded in the census, as long as the survey includes location identifiers which allow to link them to the observations included in the survey.

When the imputation methodology is applied to the construction of “literacy maps”, we replace the poverty indicator with a binary variable equal to one if the household head is literate. In this case the list of predictors used is the same list reported in Table 2, with the omission of any information on schooling. Finally, when the object of interest is a map of *municipio*-specific sanitation head counts, the

⁹In 2000, the PPP exchange rate between USD and Mexican Peso was 6.79, so that 200 Pesos correspond to approximately one PPP dollar per person per day (Heston et al. 2006).

Table 2: 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 working previous week	
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	
<i>municipio</i> -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 to predict poverty status. For the predictors of literacy or access to toilet of household head see text.

dummy variable for access to toilet is dropped from the predictors (together with its mean) and is used as dependent variable.

For each of the three economic indicators (poverty head counts, literacy and sanitation rates), we evaluate the coverage of 95% nominal confidence intervals using 250 Monte Carlo simulations. We carry out independent simulations for the three states of Chiapas, Oaxaca and Veracruz. In each replication we assume that the object of interest is a map of *municipio*-specific indicators, but that the researcher has access only to a synthetic “survey”. The survey is generated by drawing—from the state census—clusters of ten observations from 50 *municipios* selected at random without replacement.¹⁰ Because the IPUMS data set only includes a 10.6% extract of the complete micro-data, we first generate a complete “pseudo-

¹⁰All estimations are done using sampling weights, because this sampling design generate differences in the probability of selection across households.

census” with a number of observations equal to actual census population. For this purpose, we replace each observation in the extract with identical replicates in number identical to the (integer) weight provided in the data set. The pseudo-census created in this fashion is then treated as the actual (non-random) population of interest.

In each Monte Carlo replication, we first use the synthetic survey sample to estimate the parameters in the conditional probability in (3) using a logit model. These parameters are assumed to be common to all observations within the same state, and are used together with information on the predictors for all households in the state pseudo-census to estimate the indicator for each *municipio*. The confidence intervals are calculated as in (6). Finally, we calculate coverage rates as the fraction of Monte Carlo simulations where the nominal 95% confidence intervals include the “pseudo-true” *municipio*-specific indicators, calculated using the actual data included in the pseudo-census.

The results of the simulations are summarized in Figure 4. Each graph shows a histogram of the coverage rates for a state-head count pair. Each histogram is constructed using the coverage rate for a given small area (that is, a *municipio*) as an observation.¹¹

If the confidence intervals had correct coverage, the whole mass of each histogram would be centered on .95. The first conclusion one can draw from these results is that clearly this is not the case. Even if coverage rates remain acceptably large and above .75 for most small areas, all histograms show that there is a substantial fraction of *municipios* where coverage remains much below the nominal value. Coverage rates for poverty head count ratios remain below .75 in 35 percent of *municipios* in Chiapas, 38 percent in Oaxaca and 39 percent in Veracruz (see Table 3), while coverage rates are below 50% in approximately 11 percent of areas. Although the estimated confidence intervals appear to systematically overstate the precision of the estimator, they are relatively wide. For instance, the mean width of a confidence interval in the estimation of poverty head counts is .33 in Chiapas (minimum .19 and maximum .57), .38 in Oaxaca (minimum .23 and maximum .71) and .36 in Veracruz (minimum .21 and maximum .70). The second important conclusion, which is central for the argument put forth in this paper, is that in all states the distributions of coverage rates appear to be remarkably similar among different indicators. Of course, these results cannot be generalized to other contexts without scrutiny. However, they are consistent with the idea that, at least within the framework of this empirical setting, the performance of imputation methodologies to construct maps of indicators such as literacy or access to sanitation indices may be an important signal of how effectively such methodologies will be able to produce poverty maps with the correct precision. Non-income information such as schooling achievement, access to sanitation, ownership of some assets or dwelling characteristics are typically available in most censuses and they may offer, therefore, useful

¹¹Note that the histograms of the coverage rates for poverty head count are very similar but not identical to those in [Tarozzi and Deaton \(2008\)](#), because in this paper we have only included in the analysis *municipios* with at least 500 observations.

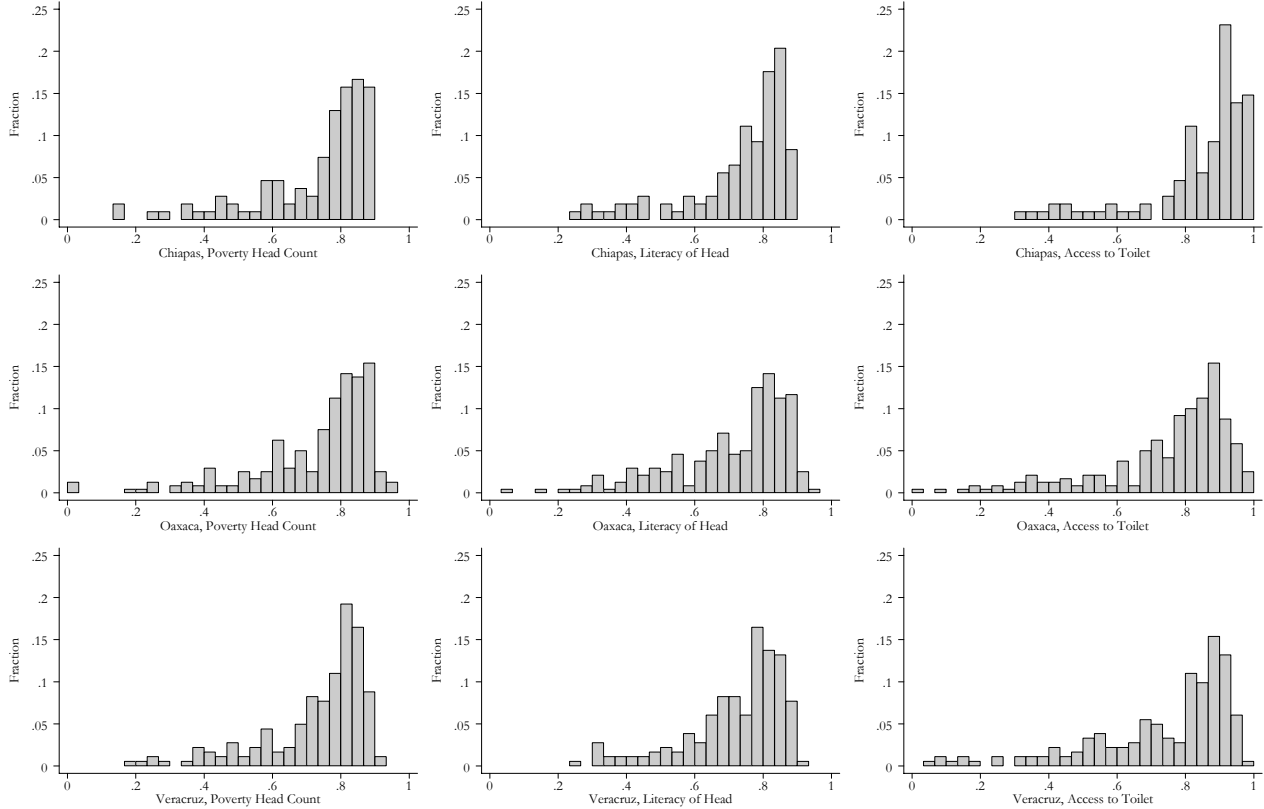


Figure 4: Histograms of coverage rates, based on 250 Monte Carlo replications.

information for researchers interested in small area estimation.

4.1 Model Selection

Our interpretation of the results presented in Figure 4 relies crucially on the fact that discrepancies between nominal and actual coverage rates are evidence of area heterogeneity. In real poverty mapping applications, discrepancies could also be due to predictors being measured differently in the census and in the survey (for instance, because of differences in how occupations or schooling are coded). However, this possibility does not arise within our simulation, because the synthetic survey data are drawn directly from the census, so comparability of predictors between the two data sources holds by construction. Another possibility is that the poor performance of mapping is due to an incorrect choice of the conditional model. On the one hand, the choice of the predictors is necessarily atheoretical, both because in a typical application the selection is limited by the variables that can be matched between the census and the survey, and because the first-stage projections are not structural relationships. On the other hand, it is conceivable that the use of a statistical criterion to select which of the predictors in Table 2 to include in the first stage may improve the performance of the mapping exercise. We experiment with two alternative selection

Table 3: Fractions of Coverage Rates below Selected Thresholds

	Poverty	Literacy	Sanitation
	<i>% municipios with coverage < .50</i>		
Chiapas	.120	.111	.074
Oaxaca	.108	.142	.121
Veracruz	.110	.093	.132
	<i>% municipios with coverage < .75</i>		
Chiapas	.352	.361	.176
Oaxaca	.375	.438	.342
Veracruz	.385	.445	.396

Source: IPUMS Mexico Census 2000. Fractions of *municipio*-specific coverage rates below .50 and .75. Each coverage rate is calculated over 250 simulations (see main text for detailed description)

procedures, one based on an information criterion and another based on the statistical significance of the first stage projection coefficients. In each of the two cases, the model is selected based on the specific synthetic survey drawn from the pseudo-census. In other words, these procedures allow the selection of a different model in each Monte Carlo replication. In both cases we also include, among the candidate predictors, a list of interactions between household-specific covariates and *municipio*-level means in Table 2. The rationale is that such interactions may help to some extent in controlling for heterogeneity in the projection coefficients across areas.

With the first selection approach, the first-stage model is chosen by using a Bayesian Information Criterion (BIC, Schwarz 1978). First, we estimate a series of univariate logit models for the dependent variable H_i where each of the K candidate predictors is entered separately. In our simulations, H_i is equal to one if the household is poor, or if it has access to sanitation, or if the head is literate. Second, we sort the models in decreasing order of pseudo- R^2 . Let $k = 1, \dots, K$ denote the index of the sorted predictors. In the third step, we calculate the BIC for K models, where each model includes the first k sorted predictors. Finally, we identify the model that maximizes the BIC. If the selected model is the j th in the ranking, all variables with a ranking not larger than j are included as predictors. Intuitively, this criterion trades off the increase in the pseudo- R^2 that results from the inclusion of more predictors with a penalization that has the purpose of limiting the number of regressors, to avoid overfitting. The resulting coverage rates are presented in Figure 5. Visual inspection is sufficient to see that the accuracy of the poverty mapping clearly declines for both poverty and literacy head counts, when compared to the results obtained with the inclusion of all predictors in Figure 4. In the estimation of poverty head counts (left-most figures), the

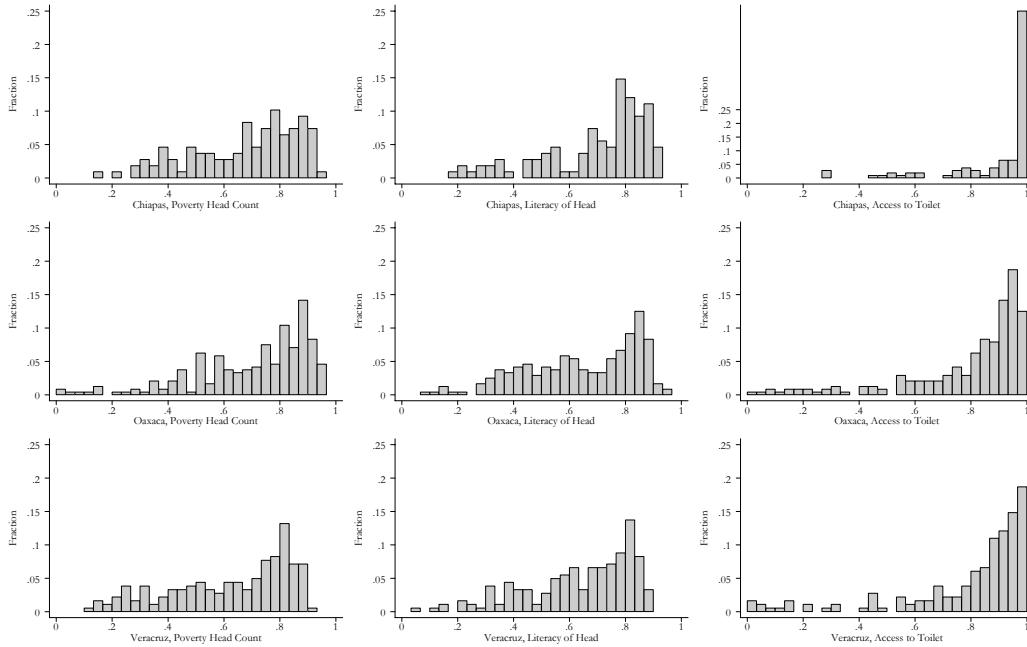


Figure 5: Histograms of coverage rates, based on 250 Monte Carlo replications, with first-stage model selected using a Bayesian Information Criterion. Note the different scale of the vertical axis in the top-right graph.

fraction of *municipios* for which coverage is below 50 percent increases from 12 to 21 percent in Chiapas, from 11 to 15 percent in Oaxaca and from 11 to 29 percent in Veracruz. In the prediction of literacy head counts, the increase is from 11 to 17 percent in Chiapas, from 14 to 26 in Oaxaca and from 9 to 23 percent in Veracruz.

The three right-most graphs in Figure 5 show instead that confidence intervals for the sanitation head counts are somewhat better than when all predictors are included, and in a substantial fraction of cases they are conservative. This is especially true for Chiapas, where coverage is larger than .95 in approximately two-thirds of the *municipios*. This finding, however, is the result of very large estimated MSEs, which reflect relatively large values of the covariance component in expression (5). In this state, the mean value of the root MSE is .17, so that on average the width of the confidence intervals is .68.

In Figure 6 we display the distribution of coverage rates estimated using a second selection criterion. In this case, the model selection proceeds by first estimating the most inclusive model, and then dropping from the prediction exercise all regressors not significant at the 10 percent level. The coverage rates in Figure 6 show that, despite this criterion leading to more parsimonious models, the distributions of coverage rates across *municipios* remain similar to those estimated when all predictors are included. The only exception is again found when we construct sanitation head counts in Chiapas, in which case confidence intervals appear to be conservative in almost half of the areas.

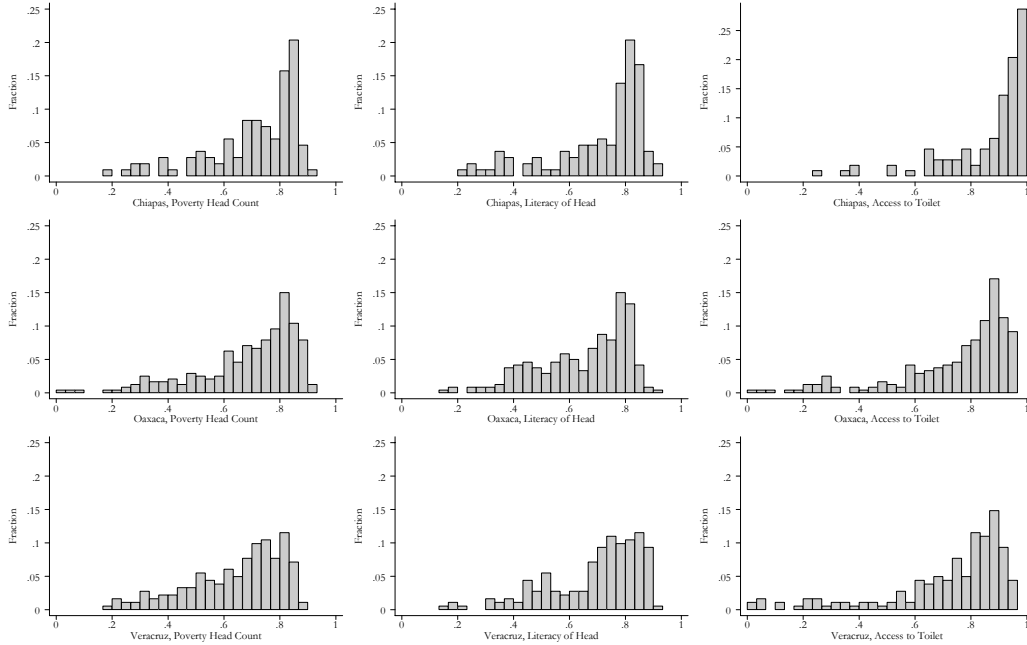


Figure 6: Histograms of coverage rates, based on 250 Monte Carlo replications, with first-stage model selected by including only predictors significant at the 10 percent level. Note the different scale of the vertical axis in the top-right graph.

Overall, these results suggest that the main cause for incorrect coverage is the presence of area heterogeneity, and not the use of an inappropriate model in the first-stage.

5 Caveats and Conclusions

Poverty mapping methodologies which are now routinely used require a substantial degree of geographic homogeneity in the relationship between income and its predictors. However, heterogeneity in local unobserved factors such as relative prices and return is likely to generate heterogeneity in such relationships, at least to some extent. The purpose of this paper is to argue that useful even if only indirect and suggestive evidence on the extent of area heterogeneity is readily available in virtually any census. Such indirect evidence is provided by non-monetary indicators such as literacy, asset ownership, work status or access to sanitation. These variables are routinely recorded in censuses, and validation exercises like those described in this paper can then be completed to gauge the extent of area heterogeneity. We argue that the same factors which would generate area heterogeneity in the projections estimated in poverty mapping would also likely generate heterogeneity, and hence incorrect coverage, in the kind of quasi-validation exercises described in Section 4.

One *caveat* to keep in mind is that the simulation described in this section cannot be seen as a true

validation. In fact, we did not have access to a true census, but rather only to a 10.6% census extract, which we have “inflated” using the sampling weights provided within the data set. It is this “pseudo-census” that all the synthetic survey samples have been drawn from, and it is the “pseudo-true” values calculated from the census abstract that we have adopted as the true values to be checked against the estimated confidence intervals. For this reason, the simulation presented here is probably best described as a Monte Carlo experiment with a Data Generating Process closely matching a real population, rather than as a true validation exercise. Of course, our suggestion is that once a researcher or statistical agency wishes to proceed with a poverty mapping exercise using real census data, *true* validations such as the ones described here be completed using information included in the census alone.

One further *caveat* is that even this form of validation exercise will be of little help if the census and the survey are completed in different years. This may be a common occurrence, because while censuses are usually decadal, surveys are in most cases carried out at shorter time intervals. But in such situation, much more than area homogeneity is required for poverty mapping to produce valid inference. Namely, one also need stability *over time* of the distribution of the predictors, because such variables are only recorded for the whole population in the census. Such assumption is clearly problematic, especially in the context of fast-growing countries where the underlying structure of the economy is likely to change rapidly. So, even if a validation such as the one described in Section 4 suggested the presence of area homogeneity in a cross-section, such evidence would be silent about the plausibility of the identifying assumptions required by poverty mapping when census and survey data have been collected in different time periods.

Finally, we emphasize again that, from a statistical point of view, the good or poor performance of a mapping methodology in constructing, say, literacy maps should *not* be used to *imply* that poverty mapping will be equally successful or unsuccessful in producing reliable inference. However, we argue that such similarity in performance should be expected. This should make the kind of exercise proposed here worthwhile in real life applications, regardless of whether one expects the assumptions underlying poverty mapping to fail as in [Tarozzi and Deaton \(2008\)](#) or to hold as in [Elbers et al. \(2008\)](#).

References

- Carroll, R., D. Ruppert, and L. Stefanski (1995). *Measurement Error in Non-Linear Models*. Chapman and Hall.
- Chen, X., H. Hong, and A. Tarozzi (2008). Semiparametric efficiency in GMM models with auxiliary data. *Annals of Statistics* 36(2), 808–843.
- 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.
- Deaton, A. and J. Muellbauer (1980, June). An Almost Ideal Demand System. *American Economic Review* 70(3), 312–326.
- Elbers, C., J. Lanjouw, and P. Lanjouw (2003). Micro-level estimation of poverty and inequality. *Econometrica* 71(1), 355–364.
- Elbers, C., P. Lanjouw, and P. G. Leite (2008). Brazil within Brazil: Testing the poverty map methodology in Minas Gerais. World Bank Policy Research Working Paper 4513.
- Fujii, T. (2007). Micro-level estimation of child undernutrition indicators and its application to targeting in Cambodia. Unpublished Manuscript, Singapore Management University.
- 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.
- Little, R. and D. Rubin (2002). *Statistical Analysis with Missing Data* (Second ed.). New York: John Wiley & Sons.
- Minnesota Population Center (2007). Integrated Public Use Microdata Series International: Version 3.0. Minneapolis: University of Minnesota.
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of The American Statistical Association* 91, 473–489.
- 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.
- Tarozzi, A. and A. Deaton (2008). Using census and survey data to estimate poverty and inequality for small areas. *Review of Economics and Statistics* (Forthcoming).
- Todd, P. E. (2007). Evaluating social programs with endogenous program placement and selection of the treated. In T. P. Schultz and J. Strauss (Eds.), *Handbook of Development Economics*, Volume IV, Chapter 60. Amsterdam: Elsevier Science.