WPS4513

POLICY RESEARCH WORKING PAPER 4513

# Brazil within Brazil:

## Testing the Poverty Map Methodology in Minas Gerais

*Chris Elbers*
*Peter Lanjouw*
*Phillippe George Leite*

The World Bank
Development Research Group
Poverty Team
February 2008

Policy Research Working Paper 4513


## Abstract

The small-area estimation technique developed for producing poverty maps has been applied in a large number of developing countries. Opportunities to formally test the validity of this approach remain rare due to lack of appropriately detailed data. This paper compares a set of predicted welfare estimates based on this methodology against their true values, in a setting where these true values are known. A recent study draws on Monte Carlo evidence to warn that the small-area estimation methodology could significantly over-state the precision of local-level estimates of poverty, if underlying assumptions of spatial homogeneity do not hold. Despite these concerns, the findings in this paper for the state of Minas Gerais, Brazil, indicate that the small-area estimation approach is able to produce estimates of welfare that line up quite closely to their true values. Although the setting considered here would seem, a priori, unlikely to meet the homogeneity conditions that have been argued to be essential for the method, confidence intervals for the poverty estimates also appear to be appropriate. However, this latter conclusion holds only after carefully controlling for community-level factors that are correlated with household level welfare.


---

This paper—a product of the Poverty Team, Development Research Group—is part of a larger effort in the department to develop tools for the analysis of poverty and income distribution. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The author may be contacted at planjouw@worldbank.org.

# Brazil within Brazil:

# Testing the Poverty Map Methodology in Minas Gerais*

Chris Ebers (Vrije University, Amsterdam)

Peter Lanjouw (World Bank)

Phillippe George Leite (World Bank)

JEL: I32, C31, C42
Key words: Small Area Statistics, Poverty, Inequality, Survey Methods, Heterogeneity

* We thank seminar participants at the World Bank Applied Micro Seminar for useful comments and suggestions. We are particularly indebted to Wasmalia Bivar, Elisa Caillaux, Antonio Dias, Mauricio Lila, Viviane Quintaes and Debora de Souza at the Instituto Brasileiro de Geografia e Estatística (IBGE) Rio de Janeiro, for making available the unit record census data for Minas Gerais for the purpose of this validation study, and for numerous comments and suggestions during the course of this study. We are also most grateful to Kathy Lindert for supporting this work and to DECRS for providing financial support (RF-P107113-RESE-BBRSB). The views in this paper are the authors' and should not be interpreted to reflect those of the World Bank or affiliated institutions.

## I. Introduction

During the past decade researchers at the World Bank and a number of partner institutions have been studying approaches to develop detailed "poverty maps" in a large number of developing countries. These maps provide estimates of (consumption or income) poverty and inequality at the local level – such as the sub-district and even community level. In general this information is not available because sample surveys do not normally permit sufficiently fine disaggregation. Yet, with ongoing efforts to apply detailed spatial targeting of public interventions, or to realize the gains from decentralization and from a greater focus on community-centered development, there is a pressing need for information on distributional outcomes at the local level. In the approach introduced in Hentschel, Lanjouw, Lanjouw and Poggi (2000) and refined further in Elbers, Lanjouw and Lanjouw (2002, 2003), household survey data are combined with unit record data from the population census in order to overcome these data constraints. The resulting welfare estimates can be used to better understand the spatial distribution of economic wellbeing and to investigate the relationship between poverty and other geographic factors.[1] These poverty maps aim to provide not only estimates of poverty or inequality levels at the local level, but to also provide a sense of the precision of these estimates. Although their potential value is well-recognized, opportunities to formally check the reliability of the local-level poverty estimates and their associated confidence intervals are rare. Such validation exercises are needed if these small-area estimation methods are to enter into regular use and their outputs are to inform policymaking.

In a recent study, Tarozzi and Deaton (2007) suggest that the methodology developed by Elbers, Lanjouw and Lanjouw (2002, 2003) - henceforth ELL (2003) – is likely to yield an overly

[1] Poverty maps have been produced or are underway in some 40-50 developing countries. Alongside their descriptive function in conveying information on the spatial distribution of poverty, the databases underpinning these maps have been used in a variety of policy-related studies (see for example, Demombynes and Ozler, 2005, Fujii and Roland Holst, 2007, Elbers et al, 2007, Araujo et al, forthcoming). In addition, the approach developed in ELL, 2003, has also prompted further methodological research aimed at, for example, estimating child nutritional outcomes at the local level (Fujii, 2005) or imputing welfare indicators across household surveys (Christiaensen and Stifel 2007, Kijima and Lanjouw, 2003)

optimistic assessment of the precision of its small area estimates. The ELL method is based on regression models of income or expenditure with random effects at the level of survey clusters. Tarozzi and Deaton (2007) argue that this methodology relies on crucial assumptions that, they claim, are likely to fail in most real settings. First, in their view, a model of income or expenditure estimated using household survey data at the level of a region, R, is unlikely to be good enough to predict welfare at the level of a small area, *A*, unless the region *R* happens to be quite homogenous. In the presence of differences in tastes and prices such an assumption of homogeneity could be contentious. Second, Tarozzi and Deaton (2007) claim that an assumption of homoskedastic and independent and identically distributed cluster random effects is very strong, because within a region, sub-regional areas are likely to be integrated. This could result in spatial correlation of residual cluster effects if regressors do not sufficiently capture such integration.

An examination of the ELL method presented in Demombynes et al (2006) provides evidence that, in contrast to the claims above, the ELL method can produce reliable welfare estimates. Demombynes et al (2006) employ data from the PROGRESA program in rural Mexico in which a population census was administered in 500 villages. This census questionnaire included a measure of household consumption amongst the variables collected. These data permit the authors to implement the ELL methodology and compare predicted welfare outcomes at the local level against actual observed values of those outcomes in the same communities. Demombynes et al (2006) demonstrate that performance of the ELL approach depends crucially on the ability to incorporate into the basic consumption or income model, locality-level explanatory variables inserted into the household survey data from outside datasets such as the census and/or other ancillary databases. The study also notes that the method does not strictly depend on an assumption of homoskedastic cluster random effects, pointing out that the simulation stage of the approach allows for a variety of assumptions as to the nature and degree of spatial correlation between clusters. Although their evidence goes some way towards discounting critics' concerns, data

limitations do prevent the Dembombynes et al (2006) study from fully addressing all doubts that have been raised.[2]

The population census of Brazil offers a second, richer, setting in which to "test" the ELL method. This dataset provides an opportunity to more completely study the applicability of the basic ELL method and its underlying assumptions. In 2000, the Brazilian Statistical Organization, henceforth IBGE, fielded two questionnaires as part of its Census data collection. The first, traditional, questionnaire was fielded to all households and includes a single-question about the income of the household head. A second questionnaire, more detailed than the traditional "short form" and with a fairly good measure of total household income, was fielded to a 12.5% sample of households within each enumeration area in the country. To date, IBGE has not published any sort of small area poverty statistics based on this household income variable because although it is fairly detailed, it is still judged to be insufficiently comprehensive.[3] While neither of the two available income measures is appealing for the purpose of producing a proper poverty map for Brazil, the databases do provide an attractive context within which to *check* the ELL methodology.

The analysis in this paper compares the predicted welfare measures obtained following application of the ELL method to the actual, observed, values. For reasons of computational ease and tractability, we focus on the single state of Minas Gerais (see further below). In addition, we focus our attention on the 12.5% census sample. From this data source, we draw synthetic household surveys mimicking the sample design of two existing surveys in Brazil: the PNAD earnings survey and the 2002/2003 POF expenditure survey. We implement the ELL methodology based on these pseudo-surveys, combining the parameter estimates from the survey-based models with unit record data from the census sample to predict poverty at the municipality level. We

[2] Notably, the Demombynes et al (2006) study is unable to confront the concerns raised regarding the impact of inter-cluster correlation on standard errors due to the fact that the target populations in this study comprised randomly assembled, non geographically-contiguous villages aggregated together into target populations of roughly 1000 households.

[3] Indeed, in the 1996 South African Census, which also includes a similar crude measure of household income, the national *income* estimate in the census amounted to only 85% of the household survey-based national *consumption* estimate, and moreover deviations between the two were not random (see Alderman et al.,2002).

compare these predictions against poverty rates calculated directly from the census sample. Our goal is to examine the accuracy of the poverty estimates and to assess whether the confidence intervals produced by the ELL method are correct. We also explore in some depth how well, and under what conditions, the income regression model – estimated at the state level – performs in capturing spatial correlation among small areas.

The paper is organized as follows. Section 2 summarizes the ELL methodology. Section 3 describes the data. Section 4 presents a set of descriptive statistics. Section 5 discusses the validation exercise and its results and section 6 concludes.

## II. The ELL Method

The ELL approach analyzes household survey data to impute consumption/income into the population census in order to generate small area welfare measures. ELL (2002, 2003), and Demombynes et al (2007) describe the methodology in detail while Tarozzi and Deaton (2007) provide a useful discussion of the method's underlying assumptions. The basic idea is straightforward. We estimate welfare measures based on a household per-capita measure of income or consumption expenditure, $y_i$. A model of $y_i$ is estimated using household survey data, with the set of explanatory variables restricted to those that are also found in, and strictly comparable to, the population census.

We regress the logarithm of $y_i$ on a set of household-level demographic, occupational and educational variables, as well as variables at the enumeration area level or some other level of aggregation above the household calculated on the basis of unit record census data (or drawn from some ancillary database):

$$\ln y_i = \mathbf{x}_i\boldsymbol{\beta} + u_i, \qquad (1)$$

where $\boldsymbol{\beta}$ is a vector of $k$ parameters and $u_i$ is a disturbance term satisfying $E[u_i|x_i] = 0$.

This 'first-stage' estimation is then carried out using nationally representative survey data, which are usually stratified at the province, or regional, level for rural and urban areas separately, but always setting household weights respecting survey's sample design.[4] This regression model of household income also allows for intra-cluster correlation in the regression residuals - failure to take account of correlation in the disturbances would result in underestimation of standard errors. Thus, the vector of disturbances, u, in (1) is decomposed as $u_{ch} = \eta_c + \varepsilon_{ch}$ where $\eta_c$ is a location component and $\varepsilon_{ch}$ a household component. (see Elbers, Lanjouw and Lanjouw, 2002, 2003 for more details). To capture latent cluster-level effects, census mean variables and other aggregate level variables are included among the set of potential regressors. The ELL method also allows for heteroskedasticity in the household-specific part of the residual, limiting the number of explanatory variables to be cautious about overfitting. Finally, the estimated variance-covariance matrix is used to obtain GLS estimates of the first-stage parameters and their variance.

The next stage is to combine the results of the first-stage regression model with census information, vector of characteristics X, to predict welfare measures and estimate prediction errors. These estimates can be generated via several routes as described in Elbers et al (2002), Pfeffermann and Tiller (2005) and Demombynes et al (2006). In brief, for each household in the census data disturbance terms, $\tilde{\eta}_c^r$ and $\tilde{\varepsilon}_{ch}^r$, are drawn from their corresponding sampling distributions estimated on the survey data. Each household in the census data then obtains an estimated expenditure, $\hat{y}_{ch}^r$, based on both predicted log expenditure, $\mathbf{x}'_{CENSUS_{ch}} \cdot \tilde{\hat{\beta}}_{SURVEY}^r$, and the disturbance terms, such that $\hat{y}_{ch}^r = \exp\left(\mathbf{x}'_{CENSUS_{ch}} \cdot \tilde{\hat{\beta}}_{SURVEY}^r + \tilde{\eta}_c^r + \tilde{\varepsilon}_{ch}^r\right)$.

Finally, this simulated expenditure is used to calculate estimates of the welfare measures for each target population. The procedure is repeated *M* times, drawing each time different set of

[4] Within each region there are usually further levels of stratification, and also clustering. At the final level, a small number of households (a cluster) are typically randomly selected from a census enumeration area.

random terms, in order to compute a point estimate (average of *M* simulations) and standard errors for each welfare measure. The ELL method has two main sources of errors in the welfare estimates: a) model error due to the fact that the parameters for the imputations are estimated; and b) idiosyncratic error associated with the fact that the actual welfare outcomes deviate from their expected value. The importance of the latter component decreases with the size of the target population. In this paper we employ the Delta Method originally presented in Elbers, Lanjouw and Lanjouw (2002) to calculate the error components.

**III. Data Source**

The main source of data in this paper is the 2000 Brazilian Population Census for the state of Minas Gerais. By construction, the 12.5% sample of the Census is representative at the level of municipalities (5,564 in Brazil as a whole in 2005). It also contains considerably more detailed information about household characteristics than the full census and the household per capita income measure includes earnings and transfers from different sources for all members of the household: main occupation; other occupations, retirement pensions, rent, other pensions, government social transfers and others.

On the basis of these data, the census sample for Minas Gerais collects information from about 606 thousand households drawn from all 853 municipalities in the state. This constitutes the target population for our study. Accordingly, we will refer to this census sample as the "census" for simplicity and all sampling and predictions will be with respect to this sub-population of Minas Gerais.

Upon our request IBGE drew 41 samples from the census file, following a sampling design that mimics that of Brazil's two main household surveys: the Household Expenditure Survey, known as POF, last collected in 2002/2003, and the Annual National Household Survey, known in

Portuguese as PNAD. These two surveys underpin most empirical work on poverty and inequality in Brazil.

The main features of our "pseudo-samples" are:

a) Using the same selected enumeration areas as from POF 2002/2003 scheme[5], IBGE selected 20 different samples with approximately 13 households per enumeration area, generating samples of 2,800 households on average. The samples here comprise about 240 enumeration areas within 151 municipalities;

b) Based on the POF sample design, IBGE selected one "pseudo-sample" from the beginning to the end. In this case, new enumeration areas were first selected, and then 13 households were selected from each enumeration areas.

c) Using the enumeration areas that had been selected for the PNAD 2005[6] survey, IBGE selected a further 20 "pseudo-samples". The PNAD sample comprises 123 municipalities and 779 enumeration areas. On average, 16 households per enumeration area were randomly selected generating samples of around 12 thousand households.

## IV. Descriptive Statistics

### 1. Why Minas Gerais?

The ELL method assumes that the coefficients $\beta$ estimated from model (1) at the level of region *R*, should be the same for each small area *A* within *R*. Tarozzi and Deaton (2007) suggest that this assumption is unappealing when the region is characterized by much heterogeneity.

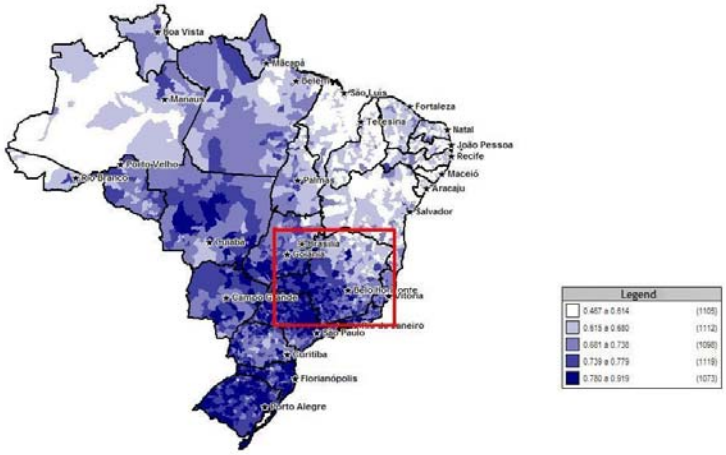
[5] One feature of the POF sample is that unlike standard techniques of sampling, IBGE stratified the state according to a combination of geographic stratum and a socio-economic status of the household head instead of only geographic stratum. Then, using systematic sampling proportional to size of each enumeration area, different enumeration areas were selected. Finally, random sampling without replacement in each enumeration area was applied to select on average 13 households per enumeration area.

[6] Unlike the POF, the PNAD sample is only based on geographic stratum, i.e., it is geographically stratified. The primary sampling units (PSU) are the municipalities, which are stratified by size (population), and selected proportional to population size. In the second stage, the enumeration areas are also selected proportional to population size (number of households). Then, a simple systematic and representative sample of households is drawn in the third stage. Note however, that the sample is not representative at the enumeration area level or at the PSU level other than those corresponding to metropolitan regions.

Differences in prices, or other sources of heterogeneity among sub-domains *A* of the region could lead to biased estimation of welfare if a unique set of parameters from a model estimated at the level of region *R* is applied. Minas Gerais state would appear to represent a setting where there are ample grounds for such concerns.

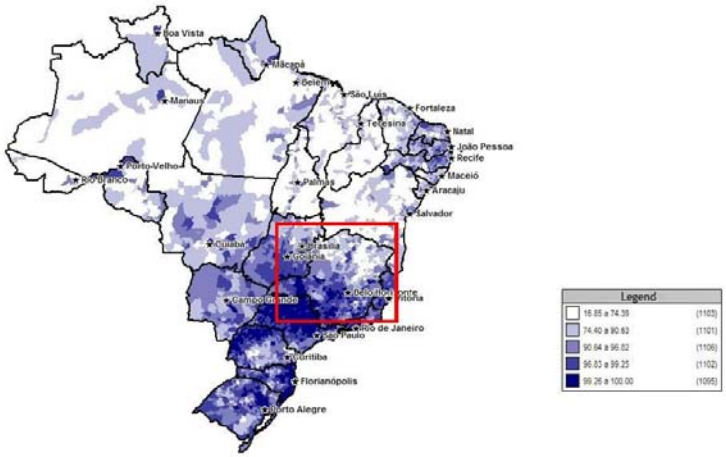
IBGE indicates that the population of the state of Minas Gerais is 17.9 million people living in 4.7 million households with a density rate of 30.5 persons per $Km^2$. The main city is Belo Horizonte, the capital, with a population of 2.2 million inhabitants living in 600 thousand households. This large and centrally located state is often referred to as "Brazil within Brazil", due to its great heterogeneity in indicators such as income, education, and infrastructure at the local level, as well as a clear regional pattern of lower welfare outcomes in the northeast of the state compared to the south. Municipalities located in the south of the state are generally well developed, while municipalities in north tend to resemble more the bordering Northeast and Center-West regions of the country. Figures 1 to 3 present different indicators for each municipality in Brazil as a whole, with Minas Gerais singled out by the red square. For all three measures, selected arbitrarily from a large set of indicators compiled in the "Atlas do Desenvolvimento Humano" by Fundação João Pinheiro and IPEA, two distinct sub regions (one dark and another light) can be distinguished – in a way that resembles the classic north-south division that applies to Brazil as a whole. Figure 1 presents the Human Development Index, HDI, in 2000. Figure 2 illustrates the share of population with access to electricity and again, the southern region indicates better access than the north. Figure 3 indicates that the south has higher average levels of adult human capital than the north.

Figure 1: Human Development Index, 2000



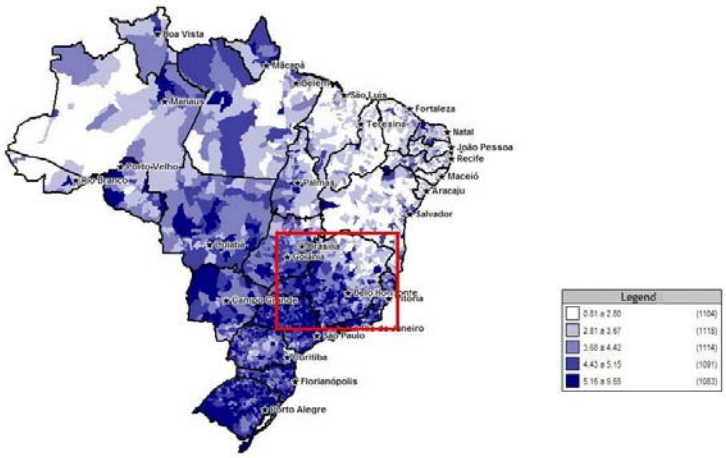

Source: Atlas do Desenvolvimento Humano 1991-2000.

Figure 2: Share of population with access to electricity, 2000




Source: Atlas de Desenvolvimento Humano 1991-2000.

Figure 3: Average years of schooling of adult population, 2000




Source: Atlas do Desenvolvimento Humano 1991-2000

Figures 4 to 8 focus specifically on Minas Gerais and highlight further heterogeneity within the state with 5 additional indicators of development at the municipality level: household per capita income, infant mortality rates, share of children aged 7 to 14 enrolled in primary schools, life expectancy at birth and an index of longevity. Irrespective of indicator, the south always contrasts with the north. These figures indicate that spatial correlation of welfare is at least partly captured by *observable* household and location characteristics. Consequently, it is not obvious that such heterogeneity would result in large spatial correlation of *unobservable* location effects as well (as seems to be one of Tarozzi and Deaton's concerns).

Figure 4: Household per capita income, 2000

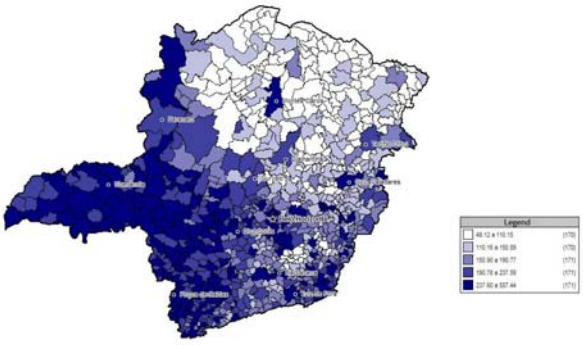
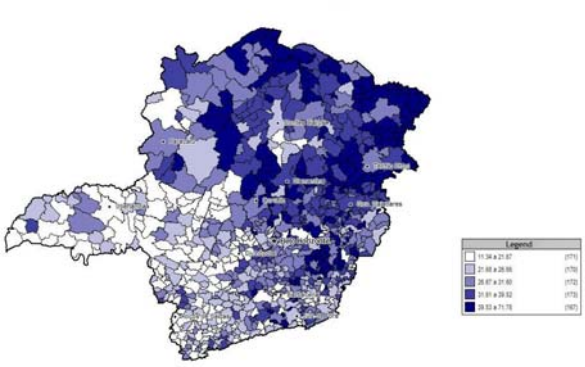


Figure 5: Infant mortality rate, 2000



Source: Atlas do Desenvolvimento Humano 1991-2000.

Figure 6: Share of children aged 7 to 14 enrolled on primary schools, 2000

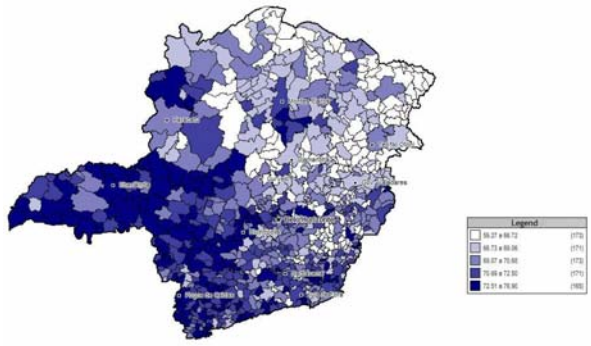
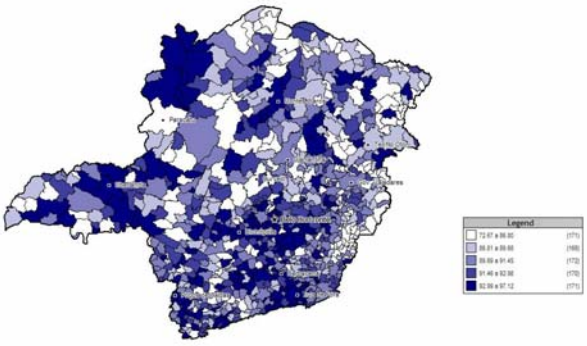
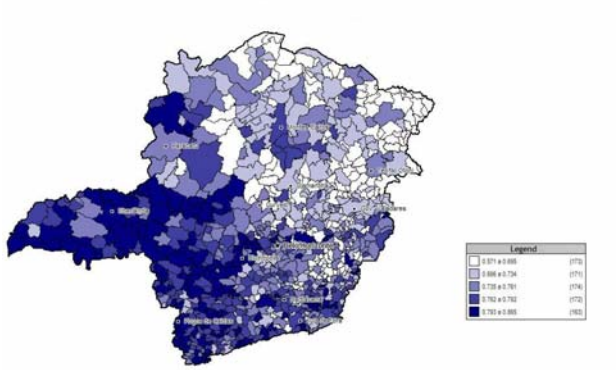


Figure 7: Life expectancy, 2000



Source: Atlas do Desenvolvimento Humano 1991-2000.

Figure 8: Human development index of longevity, 2000

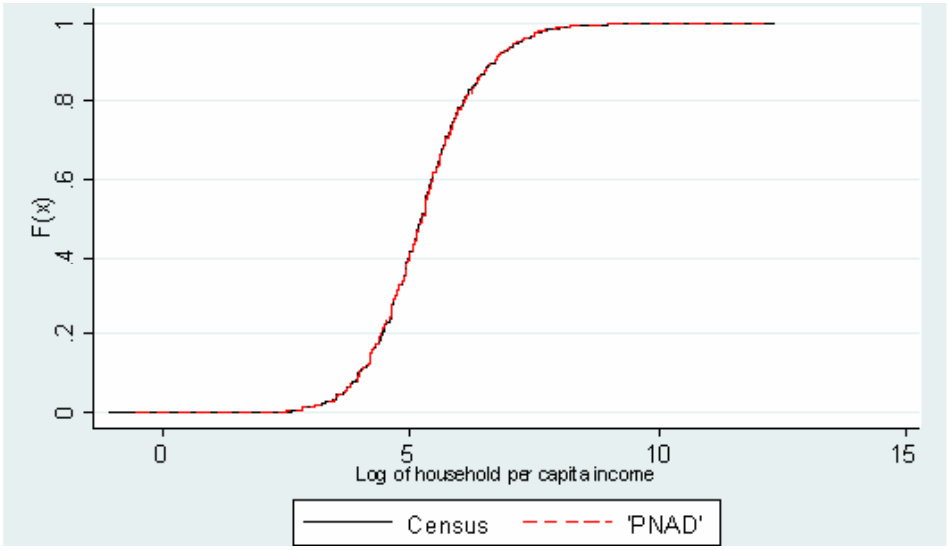


Source: Atlas do Desenvolvimento Humano 1991-2000.

## 2. Sample quality

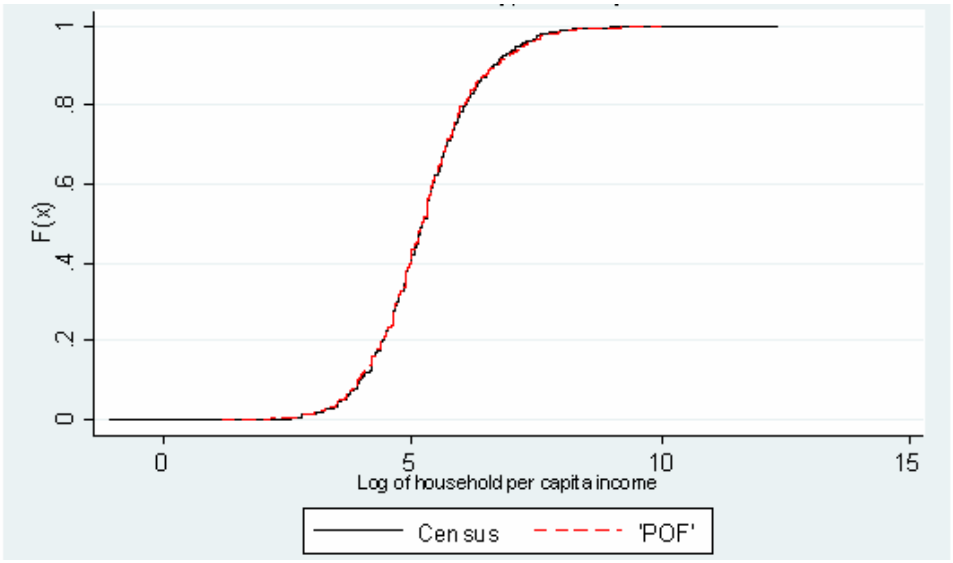
To assess the quality of the "pseudo-samples", we compare the estimated household per capita income distribution in the samples against that from the census. Figure 9 presents the cumulative density function of one of the PNAD-type samples while figure 10 considers one of the POF-type samples. Both figures indicate that surveys closely replicate the observed distribution of the census.

Figure 9: Cumulative density function of the logarithm of per capita income – PNAD sample and Census




Source: Census; Pseudo-survey and Authors' calculation.

Figure 10: Cumulative density function of the logarithm of per capita income – POF sample and Census




Source: Census; Pseudo-survey and Authors' calculation.

We next compare a set of potential regressors from the "pseudo-samples" and the census to confirm that both data sources have variables with a similar distribution that can potentially be used later in stage 1. Tables 1 and 2 below indicate that point estimates are close and that statistical tests would fail to reject the null hypothesis that the distributions were equal at 5% level of confidence.

Table1: Comparing Census and PNAD pseudo-sample variables

| | Census | | 'PNAD' | | Δ |
|---|---|---|---|---|---|
| | μ | σ | μ | σ | |
| Washing Machine | 0.27 | 0.00 | 0.28 | 0.00 | 0.98 |
| Paved street | 0.68 | 0.00 | 0.68 | 0.00 | 0.99 |
| Male head | 0.79 | 0.00 | 0.78 | 0.00 | 1.01 |
| TV set | 0.90 | 0.00 | 0.90 | 0.00 | 1.00 |
| Head with no schooling | 0.14 | 0.00 | 0.14 | 0.00 | 1.01 |
| Head with 1-3 y.s. | 0.21 | 0.00 | 0.21 | 0.00 | 1.00 |
| Head with 4-7 y.s. | 0.36 | 0.00 | 0.36 | 0.01 | 1.00 |
| Head with 8-10 y.s. | 0.11 | 0.00 | 0.11 | 0.00 | 0.99 |
| Head with 11-14 y.s. | 0.12 | 0.00 | 0.12 | 0.00 | 0.99 |
| Family type 1 | 0.03 | 0.00 | 0.02 | 0.00 | 1.03 |
| Family type 2 | 0.07 | 0.00 | 0.07 | 0.00 | 1.02 |
| Share of Inactive people at the district level | 0.14 | 0.00 | 0.15 | 0.00 | 0.96 |
| Age of the Head | 46.44 | 0.03 | 46.73 | 0.14 | 0.99 |
| Number of Chidren in the household | 0.64 | 0.00 | 0.64 | 0.01 | 1.00 |
| Metropolitan Region | 0.24 | 0.00 | 0.24 | 0.00 | 1.01 |
| House | 0.93 | 0.00 | 0.93 | 0.00 | 1.00 |
| Apartment | 0.07 | 0.00 | 0.06 | 0.00 | 1.07 |
| Average Income of the head at the municipio level | 617.92 | 0.89 | 620.07 | 1.42 | 1.00 |

Source: Census; Pseudo-survey and Authors' calculation. A standard error of 0.00 indicates a level below 0.005.

Table2: Comparing Census and POF pseudo-sample variables

| | Census | | 'POF' | | Δ |
|---|---|---|---|---|---|
| | μ | σ | μ | σ | |
| Piped water connection | 0.82 | 0.00 | 0.84 | 0.01 | 0.99 |
| Well in the property | 0.15 | 0.00 | 0.14 | 0.01 | 1.06 |
| Washing Machine | 0.27 | 0.00 | 0.27 | 0.01 | 1.02 |
| White head | 0.52 | 0.00 | 0.53 | 0.01 | 0.98 |
| Male head | 0.79 | 0.00 | 0.78 | 0.01 | 1.01 |
| TV set | 0.90 | 0.00 | 0.90 | 0.01 | 0.99 |
| Head with no schooling | 0.14 | 0.00 | 0.14 | 0.01 | 1.04 |
| Head with 1-3 y.s. | 0.21 | 0.00 | 0.23 | 0.01 | 0.94 |
| Head with 4-7 y.s. | 0.36 | 0.00 | 0.35 | 0.01 | 1.03 |
| Head with 8-10 y.s. | 0.11 | 0.00 | 0.11 | 0.01 | 0.96 |
| Head with 11-14 y.s. | 0.12 | 0.00 | 0.12 | 0.01 | 1.01 |
| Family type 1 | 0.03 | 0.00 | 0.03 | 0.00 | 0.89 |
| Family type 2 | 0.07 | 0.00 | 0.07 | 0.00 | 1.00 |
| Share of Inactive people at the district level | 0.43 | 0.00 | 0.43 | 0.00 | 0.99 |
| Share of Informal workers at the district level | 0.32 | 0.00 | 0.32 | 0.00 | 0.99 |
| Share of Formal workers at the district level | 0.29 | 0.00 | 0.28 | 0.00 | 1.03 |
| Age of the Head | 46.44 | 0.03 | 47.03 | 0.32 | 0.99 |
| Number of Chidren in the household | 0.64 | 0.00 | 0.65 | 0.03 | 0.98 |
| Ratio of rooms serving as dorms over total number of rooms | 0.68 | 0.00 | 0.68 | 0.01 | 0.99 |
| Average Household Per Capita Income at the district level | 1014.50 | 2.29 | 1009.19 | 8.59 | 1.01 |

Source: Census; Pseudo-survey and Authors' calculation. A standard error of 0.00 indicates a level below 0.005.

## V. Validation Exercise

### 1. Spatial differences in returns

As noted above, the ELL method is predicated on the assumption that a model estimated for household survey data at an aggregated region level, *R*, can generate predictors and error term distributions that can be used to estimate welfare in a small area *A*. Ideally, one would like to estimate a separate model for each area, *A*, or at least allow for different slopes for different areas within a single model. However, no sample survey is representative at the small area level, or even covers all small areas. Thus the ELL method cannot be based on separate models for each small area and it is impossible to control for small area effects via a fixed-effects specification. Instead, the ELL approach inserts into the survey a number of variables aggregated at the small-area level,

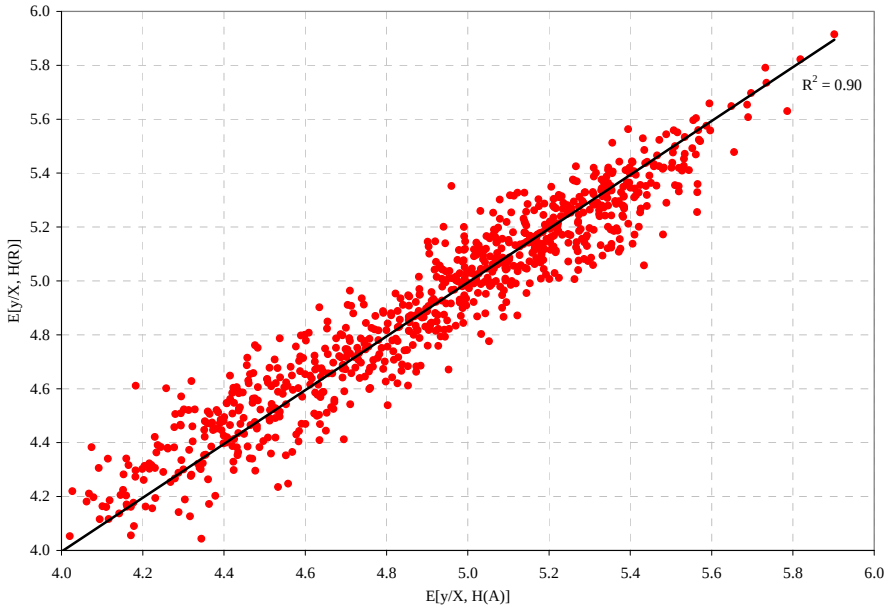
calculated from the census, or obtained from ancillary data sources. Some of these are then included in the model specification to capture small area heterogeneity.[7]

To probe the method's success in this respect we use census data to estimate two sets of models. The first set comprises a single state-level model (including a number of municipality level aggregates as regressors). The second comprises a set of 853 municipality-specific models. We compute the average predicted value of income for each municipality on the basis of the two sets of models. Figure 11 shows that municipality-level predictions are closely centered on the main diagonal indicating a close match of the pairs. In other words, predicted income is not markedly different if the model is estimated with our state-level model as opposed to a municipality-specific model. The estimated correlation among the two predictions is 0.90, and 80% of the conditional means based on the municipality-specific model are found within the 95% confidence interval estimated using the state-level model. It seems that specifying a model at the level of region *R* is not particularly problematic for estimating welfare at area *A* as long as the model captures local heterogeneity.[8]

[7] Enumeration area, district and municipality level variables, such as total population, formal sector employment shares, literacy rates, availability of publicly provided water and sanitation services, and so on, are generally found to be strongly correlated with household per capita income, even after controlling for household level characteristics. To assess how well our strategy of using such variables works we can estimate models for of our set of samples using first a fixed effects specification and then using local-level averages instead. We observe a ratio of the two R-squares in the range of of 0.95-1.00, confirming that the latter model performs nearly as well as a fixed-effects specification (see also Demombynes et al , 2006). For further detail on the models estimated, see further below and also Appendix 1.

[8] By removing all municipal-level averages from the state-level model, the correlation in the sample decreases to 0.50 indicating that small areas heterogeneity control must be taken into account during model specification. This heterogeneity affects the precision of estimations and can also lead to an overestimation of the error component.

Figure 11: Conditional independence assumption test - Comparing expected values at municipality level A estimated through a single model at state level R and multiple models estimated at A level




Source: Census and Authors' calculation

## 2. Inter-cluster correlation of errors

We next ask whether a good model specification in (1) addresses concerns about inter-cluster correlation of regression residuals. In principle, there can be many levels at which a location effect occurs. To see how such inter-cluster correlation matters, suppose we expand on (1) and consider the following:

$\log y_{ach} - \log \hat{y}_{\eta ch} = u_{ach} = \eta_a + e_{ac} + \varepsilon_{ach}$, where $\log \hat{y}_{ach} = \mathbf{x}'_{CENSUS_{ach}} \cdot \tilde{\hat{\beta}}^{r}_{SURVEY}$.

This specification allows for a separate 'area level effect- η' (e.g at the municipality level), and a 'cluster level effect - *e*' (enumeration area), alongside an idiosyncratic household level effect, ε. The inter-cluster correlation coefficient and inter-area correlation coefficient can then be estimated, respectively, by $\rho_c = \frac{\left(\sigma_\eta^2 + \sigma_e^2\right)}{\sigma_u^2}$ and $\rho_a = \frac{\sigma_\eta^2}{\sigma_u^2}$. These separate components are quite important for the simulation phase of the ELL method. As emphasized by Tarozzi and Deaton (2007), the variance of the welfare predictions in the final phase can be understated when inter-

cluster and inter-area correlations are large and are not explicitly accounted for. Central to the ELL approach is the fact that it is not generally possible to separate the overall location effect into the area level effect 'η' and the cluster level effect '*e*' and, in general, just a single location effect can be calculated. Thus, in the simulation phase, the ELL method requires that one either assumes that the estimated location effect measured by $\sigma_{\eta}^{*2} = \left(\sigma_{\eta}^{2} + \sigma_{e}^{2}\right)$ is entirely a cluster level effect - an optimistic assumption that rules out any correlation at a higher level - or that it occurs entirely at the area level, a conservative assumption that will likely lead to an overstatement of the variance of the estimate.[9]

How large are inter-cluster correlations in practice? Given the availability of income data in the Census, we can analyze in this study the presence of inter-cluster correlation at multiple levels. Within the state, we discern five possible locational levels above the household at which inter-cluster correlations might apply: the meso-region, micro-region, municipality, district and enumeration level (see Table 3). We use the full census data to estimate a single model for the state as a whole, including a set of locational controls (aggregated from unit record census data and included as regressors) at the enumeration area, district and municipality level. We apply mixed-effects maximum likelihood estimation and decompose the overall error term into a household component and separate sub-components for each of the five respective locational levels described above. Tables 4a and 4b indicate that irrespective of the number of areas one allows for, the bulk of the overall location-effect arises at the enumeration area level. While some contribution does derive from correlations at a higher level, they account for less than 0.5% of the total variance. Moreover, after having controlled for locational characteristics in our specification, the entire inter-cluster correlation contribution (including the EA-level effect) remains below 3% of the total

[9] See details in Elbers, Lanjouw and Lanjouw (2002) or Demombynes et al (2006). Tarozzi and Deaton (2007) claim that application of the conservative assumption, while feasible, is not appealing as it would inevitably result in estimates of poverty that are so imprecise as to be unusable. For this reason they conclude that the "optimistic" assumption is essentially unavoidable for the ELL methodology. We shall empirically assess this assertion in greater detail below.

variance.[10] This latter percentage rises to about 7-8% if a "naïve" model that does not control for locational effects is estimated.[11] In addition, with the naïve model the importance of locational effects at the higher level become more pronounced (Table 4c).

Table 3: Breakdown of localities in Minas Gerais

| Locality-type | No. of Localities | Number of Households Per Locality | | |
|---|---|---|---|---|
| | | Minimum | Average | Maximum |
| Meso-region | 12 | 10519 | 48637 | 158153 |
| Micro-region | 66 | 1601 | 88430 | 114589 |
| Municipality | 853 | 46 | 684 | 61852 |
| District | 1.568 | 3 | 372 | 39410 |
| Enumeration area | 22,211 | 1 | 26 | 153 |

Source: Census and authors' calculation.

Table 4a: Inter-cluster variances: Preferred Model

| Locality-type | M1 | M2 | M3 | M4 | M5 | M6 | M7 |
|---|---|---|---|---|---|---|---|
| Meso-Region | 0.00001 | 0.00003 | - | - | - | - | - |
| Micro-Region | 0.00007 | 0.00006 | - | - | - | - | - |
| Municipality | 0.00296 | 0.00178 | 0.00408 | - | - | - | 0.00187 |
| District | - | 0.00229 | - | 0.00630 | - | 0.00415 | 0.00223 |
| E.A. | 0.01042 | 0.00948 | - | - | 0.01190 | 0.00948 | 0.00948 |
| Household | 0.42292 | 0.42315 | 0.43252 | 0.43173 | 0.42315 | 0.42316 | 0.42316 |

Source: Census and authors' calculation. Models M1 to M7 represent models with different nested error structures.

Table 4b: Percentage contribution to total variance: Preferred Model

| Locality-type | M1 | M2 | M3 | M4 | M5 | M6 | M7 |
|---|---|---|---|---|---|---|---|
| Meso-Region | 0.0 | 0.0 | - | - | - | - | - |
| Micro-Region | 0.0 | 0.0 | - | - | - | - | - |
| Municipality | 0.7 | 0.4 | 0.9 | - | - | - | 0.4 |
| District | - | 0.5 | - | 1.4 | - | 0.9 | 0.5 |
| E.A. | 2.4 | 2.2 | - | - | 2.7 | 2.2 | 2.2 |
| Household | 96.9 | 96.9 | 99.1 | 98.6 | 97.3 | 96.9 | 96.9 |

Source: Census and author's calculation. Models M1 to M7 represent models with different nested error structures.

Table 4c: Percentage contribution to total variance: Naive Model

| Locality-type | M1 | M2 | M3 | M4 | M5 | M6 | M7 |
|---|---|---|---|---|---|---|---|
| Meso-Region | 0.98 | 1.08 | - | - | - | - | |
| Micro-Region | 1.02 | 1.04 | - | - | - | - | |
| Municipality | 1.48 | 0.91 | 3.63 | | - | - | 2.78 |
| District | - | 0.99 | - | 4.77 | - | 4.10 | 1.14 |
| E.A. | 3.48 | 3.25 | - | - | 5.96 | 3.23 | 3.24 |
| Household | 93.04 | 92.73 | 96.37 | 95.23 | 94.04 | 92.68 | 92.83 |

Source: Census and author's calculation. Models M1 to M7 represent models with different nested error structures.

[10] Note while it is not typically possible to examine the contribution of multiple higher-level location effects absent the availability of the kind of data we use here, conventional surveys do sometimes employ a multi-clustered sampling design. In such cases one can carry out a similar investigation of the separate contribution of the EA-level effect relative to a single higher, "district" or "municipality" level effect. Experience with poverty mapping applications in other settings suggest that the pattern observed here, of an overwhelming share deriving from the EA level, is quite general.

[11] See also further below.

**3. Spatial analysis**

It is also of interest to directly analyze the spatial correlation of the error term generated by the model specification using our pseudo-survey data. To this end, we estimate the model for the state on the basis of one (arbitrarily selected) survey and then compare in the census data, the actual household average per capita income against predicted income generated by: $\hat{\log y}_{ach} = \mathbf{x}'_{CENSUS_{ach}} \cdot \tilde{\tilde{\beta}}^{r}_{SURVEY}$. A spatial correlation test can be defined by the following:

a) Let $\overline{\log y}_c = \alpha_0 + \alpha_1 \cdot \overline{\hat{\log y}}_c + \alpha_2 \cdot W\left[\overline{\log y}_j - \overline{\hat{\log y}}_j\right] + \varepsilon_c$ represent an equation at the enumeration area level *c* of the average log per capita consumption, $\overline{\log y}_c$, the average of the predicted log consumption estimated using parameters estimated from the sample, $\overline{\hat{\log y}}_c = \overline{\mathbf{x}'_{CENSUS_c}} \cdot \tilde{\tilde{\beta}}^{r}_{SURVEY}$; and the spatial error component not explained by the model computed on the basis of a local weight matrix W that is derived from contiguity with, or distance from, geographical areas *j* that neighbor enumeration area *c* (see below). [12]

b) If the model doesn't capture local conditions well, we would expect significant parameter estimates for all three parameters. On the other hand, if the model is well specified, we would expect to find $\hat{\alpha}_0 = 0$, $\hat{\alpha}_1 = 1$ and $\hat{\alpha}_2 = 0$.

Our regression results confirm that EA controls go a considerable way towards removing spatial correlation. Table 5 shows that with no locational controls (Model A), coefficients are always significant suggesting that spatial correlation is present. The null hypothesis test

[12] Spatial weights refer to the way in which we define neighboring. Rook and Queen Contiguity spatial weights use two different definitions of common boundaries to define neighboring. This sort of weighting matrix need not to be limited to first order contiguity; higher order contiguity boundaries can be set using the algorithm by Anselin and Smirnov (1996). Rook and Queen Contiguity spatial weighting often leads to a very unbalanced structure. Larger units can have more neighbors and small units a smaller number of neighbors. The solution is to set a unique number of neighbors for all areas by creating a k-nearest neighbor weighting matrix. When geo-referenced coordinates are available, the spatial weights can be derived from the distance between different points. Euclidean distance weighting fixes a specified distance and then counts the number of neighbors that fall within that distance. In this paper we fix the distance to 2.1 km, unless otherwise specified.

$\hat{\alpha}_0 = \hat{\alpha}_2 = 0$ and $\hat{\alpha}_1 = 1$ is rejected at any level of confidence. Once we move to model B, which includes the EA-level aggregated variables, the intercept becomes insignificant for all weighting schemes and the predicted income is statistically equal to 1 in all specifications. The spatial correlation coefficient is still generally significant although not necessarily at all significance levels and for all weighting schemes. The null hypothesis of no spatial correlation is still rejected at all levels of significance with the 5 and 10 neighbor weighting schemes, but fails to be rejected at the 1% significance level for the Euclidian distance weighting scheme.[13]

Table 5 - Spatial regression of the observed value on predicted and error component to measure any remaining spatial correlation.

| | Model A: No local controls | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Weighting Matrix** | **5-Neighbors** | | | **10-Neighbors** | | | **Euclidian distance** | | |
| | Coefficient | s.e. | | Coefficient | s.e. | | Coefficient | s.e. | |
| $\alpha_0$ | -0.748 | 0.109 | ** | -0.662 | 0.095 | ** | -1.145 | 0.174 | ** |
| $\alpha_1$[1] | 1.135 | 0.019 | ** | 1.118 | 0.016 | ** | 1.209 | 0.032 | ** |
| $\alpha_2$ | 0.406 | 0.029 | ** | 0.514 | 0.022 | ** | 0.005 | 0.002 | * |
| | Model B: With local controls | | | | | | | | |
| **Weighting Matrix** | **5-Neighbors** | | | **10-Neighbors** | | | **Euclidian distance** | | |
| | Coefficient | s.e. | | Coefficient | s.e. | | Coefficient | s.e. | |
| $\alpha_0$ | 0.137 | 0.092 | | 0.130 | 0.096 | | 0.382 | 0.233 | |
| $\alpha_1$[1] | 0.976 | 0.015 | | 0.978 | 0.016 | | 0.929 | 0.040 | |
| $\alpha_2$ | 0.546 | 0.055 | ** | 0.650 | 0.030 | ** | 0.014 | 0.005 | * |

Source: Authors' calculation.

Note: ** Significant at 1%;*significant at 5%

[1] Test whether coefficient is equal to 1; ** means we do have evidence to reject that $\alpha = 1$.

To summarize, the addition of local variables significantly diminishes enumeration area correlation in the deviation of the local welfare measure from its prediction. It is important to emphasize that the ELL approach depends on a model specification that is carefully chosen from a set of "matched" variables between the survey and census and that includes, in addition, EA, district, municipality and/or other aggregated level variables in order to reduce or capture effects of integrated small areas. Note, however, that applicability of the ELL method does not hinge on fully independent prediction errors at the enumeration area level. The approach may do quite well as

[13] See Appendix 2 for an alternative perspective on local geographic correlation of the unexplained component of simulated income.

long as unobserved location effects are sufficiently small. In the presence of some remaining correlation, and no direct information as to which specific level the area effect pertains, the conservative stance is to take the observed correlation of the deviation from predictions within enumeration areas and apply these at the level of the "target-population" (the level at which estimates of poverty will be calculated in the census) when carrying out the simulations with the census data. We demonstrate the impact of this strategy in the next section.

## 4. Implementing the ELL approach

For each of our 41 pseudo-samples, we run OLS regressions and obtain an R-square ranging from 51.6% to 62.4%. In Appendix 1 we present two examples: one corresponding to a POF-type pseudo sample and the other corresponding to a PNAD-type sample. The model specifications have at least 17 variables and the largest one has as many as 45 variables including the locality-level aggregates. For this exercise, we define the municipality as the cluster for the POF sample-type surveys because of a nearly 1 to 1 matching of municipalities and enumeration areas. For the PNAD-type sample, we set the enumeration areas in a given municipality as a single cluster.[14] Table 6 below indicates a location effect ranging from 2% to 4.1% which is relatively small, and in line with what has been observed in other applications. The location effect is 50%-100% larger when the model is estimated without location controls.

[14] In the case of the POF type sample, we observe only 240 enumeration areas selected in 151 municipalities while in PNAD sample, 779 enumeration areas in 123 municipalities.

Table 6: Location Effect and $R^2$ estimated on the basis of 41 pseudo surveys

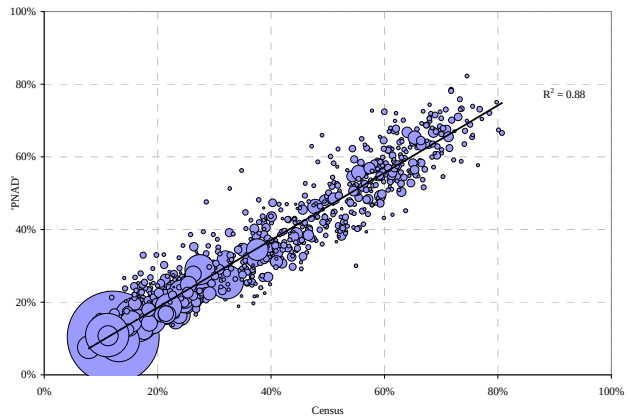
| | | Without local Controls: Naïve | | | | With local Controls: Preferred | | | |
|---|---|---|---|---|---|---|---|---|---|
| Pseudo-Sample | | $R^2$ | $\sigma_u$ | $\sigma^*_c$ | $\sigma^{*2}_c / \sigma^2_u$ | $R^2$ | $\sigma_u$ | $\sigma^*_c$ | $\sigma^{*2}_c / \sigma^2_u$ |
| 'PNAD' obs: 11,721 | 1 | 0.538 | 0.740 | 0.179 | 0.0584 | 0.602 | 0.686 | 0.101 | 0.0215 |
| 'PNAD' | 2 | 0.536 | 0.742 | 0.181 | 0.0597 | 0.590 | 0.696 | 0.106 | 0.0233 |
| 'PNAD' | 3 | 0.583 | 0.707 | 0.150 | 0.0450 | 0.596 | 0.696 | 0.098 | 0.0199 |
| 'PNAD' | 4 | 0.568 | 0.704 | 0.175 | 0.0619 | 0.593 | 0.684 | 0.094 | 0.0189 |
| 'PNAD' | 5 | 0.580 | 0.691 | 0.167 | 0.0587 | 0.602 | 0.674 | 0.101 | 0.0224 |
| 'PNAD' | 6 | 0.531 | 0.740 | 0.204 | 0.0762 | 0.568 | 0.710 | 0.107 | 0.0227 |
| 'PNAD' | 7 | 0.529 | 0.732 | 0.199 | 0.0736 | 0.567 | 0.703 | 0.097 | 0.0192 |
| 'PNAD' | 8 | 0.540 | 0.726 | 0.190 | 0.0686 | 0.573 | 0.699 | 0.102 | 0.0212 |
| 'PNAD' | 9 | 0.551 | 0.715 | 0.186 | 0.0678 | 0.578 | 0.693 | 0.114 | 0.0273 |
| 'PNAD' | 10 | 0.534 | 0.727 | 0.187 | 0.0659 | 0.560 | 0.708 | 0.119 | 0.0284 |
| 'PNAD' | 11 | 0.447 | 0.800 | 0.231 | 0.0837 | 0.506 | 0.756 | 0.107 | 0.0200 |
| 'PNAD' | 12 | 0.525 | 0.744 | 0.190 | 0.0652 | 0.554 | 0.722 | 0.115 | 0.0253 |
| 'PNAD' | 13 | 0.563 | 0.708 | 0.188 | 0.0706 | 0.586 | 0.690 | 0.129 | 0.0348 |
| 'PNAD' | 14 | 0.531 | 0.733 | 0.190 | 0.0670 | 0.556 | 0.713 | 0.112 | 0.0249 |
| 'PNAD' | 15 | 0.522 | 0.736 | 0.201 | 0.0748 | 0.555 | 0.710 | 0.110 | 0.0240 |
| 'PNAD' | 16 | 0.554 | 0.715 | 0.175 | 0.0601 | 0.575 | 0.698 | 0.114 | 0.0268 |
| 'PNAD' | 17 | 0.554 | 0.715 | 0.180 | 0.0630 | 0.576 | 0.698 | 0.126 | 0.0326 |
| 'PNAD' | 18 | 0.529 | 0.734 | 0.205 | 0.0778 | 0.559 | 0.710 | 0.130 | 0.0333 |
| 'PNAD' | 19 | 0.555 | 0.716 | 0.175 | 0.0598 | 0.572 | 0.702 | 0.127 | 0.0324 |
| 'PNAD' | 20 | 0.529 | 0.737 | 0.192 | 0.0677 | 0.553 | 0.716 | 0.116 | 0.0263 |
| 'POF' obs: 2,800 | 1 | 0.571 | 0.725 | 0.144 | 0.0393 | 0.579 | 0.718 | 0.123 | 0.0293 |
| 'POF' | 2 | 0.582 | 0.699 | 0.152 | 0.0471 | 0.590 | 0.693 | 0.131 | 0.0355 |
| 'POF' | 3 | 0.568 | 0.710 | 0.150 | 0.0445 | 0.581 | 0.700 | 0.112 | 0.0257 |
| 'POF' | 4 | 0.578 | 0.696 | 0.153 | 0.0483 | 0.587 | 0.689 | 0.126 | 0.0335 |
| 'POF' | 5 | 0.579 | 0.706 | 0.136 | 0.0371 | 0.584 | 0.702 | 0.125 | 0.0317 |
| 'POF' | 6 | 0.591 | 0.700 | 0.160 | 0.0525 | 0.599 | 0.693 | 0.135 | 0.0381 |
| 'POF' | 7 | 0.590 | 0.698 | 0.145 | 0.0429 | 0.606 | 0.688 | 0.116 | 0.0285 |
| 'POF' | 8 | 0.576 | 0.707 | 0.161 | 0.0518 | 0.587 | 0.698 | 0.133 | 0.0364 |
| 'POF' | 9 | 0.579 | 0.702 | 0.156 | 0.0491 | 0.589 | 0.693 | 0.128 | 0.0341 |
| 'POF' | 10 | 0.603 | 0.691 | 0.137 | 0.0394 | 0.612 | 0.684 | 0.103 | 0.0228 |
| 'POF' | 11 | 0.583 | 0.712 | 0.143 | 0.0406 | 0.593 | 0.704 | 0.117 | 0.0278 |
| 'POF' | 12 | 0.582 | 0.697 | 0.149 | 0.0454 | 0.595 | 0.687 | 0.110 | 0.0255 |
| 'POF' | 13 | 0.595 | 0.689 | 0.154 | 0.0502 | 0.608 | 0.677 | 0.113 | 0.0279 |
| 'POF' | 14 | 0.598 | 0.694 | 0.133 | 0.0365 | 0.607 | 0.687 | 0.104 | 0.0229 |
| 'POF' | 15 | 0.585 | 0.695 | 0.155 | 0.0495 | 0.594 | 0.687 | 0.127 | 0.0339 |
| 'POF' | 16 | 0.599 | 0.686 | 0.164 | 0.0569 | 0.614 | 0.674 | 0.114 | 0.0284 |
| 'POF' | 17 | 0.598 | 0.692 | 0.135 | 0.0378 | 0.604 | 0.687 | 0.110 | 0.0256 |
| 'POF' | 18 | 0.623 | 0.664 | 0.141 | 0.0451 | 0.631 | 0.657 | 0.118 | 0.0323 |
| 'POF' | 19 | 0.616 | 0.684 | 0.136 | 0.0395 | 0.625 | 0.676 | 0.102 | 0.0227 |
| 'POF' | 20 | 0.625 | 0.668 | 0.148 | 0.0490 | 0.625 | 0.659 | 0.111 | 0.0282 |
| 'POF' | 21 | 0.541 | 0.744 | 0.168 | 0.0508 | 0.558 | 0.732 | 0.108 | 0.0219 |

Source: Authors' Calculation.

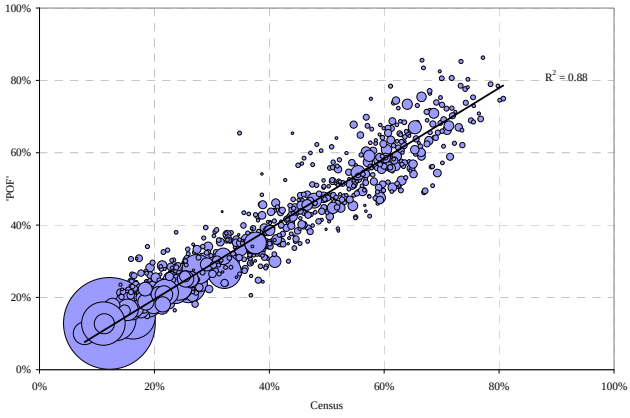
Following the model estimation stage, we apply the cluster effect at the municipality level in the simulations, generating FGT(α) measures for each one of 853 municipalities in the state of Minas Gerais. Plotting on the horizontal axis the observed FGT(α) measured on the basis of the Census data and in the vertical axis the simulated FGT(α) measure, we can examine whether predictions are close to the main diagonal. Figures 12 to 14 show the correlation among observed and simulated FGT measures for two out of 41 poverty maps we have produced. The size of the bubbles represents the size of the municipality to illustrate some imprecision of simulations in small municipalities. The figures demonstrate that poverty estimates are randomly assigned around the main diagonal and the $R^2$ representing the correlation of between observed and estimated poverty rates ranges from 75% to 90% depending on the survey type and FGT(α) measure.

Figure 12 - FGT(0) measures at Municipality level, ELL simulations and True Census estimates

PNAD pseudo-sample



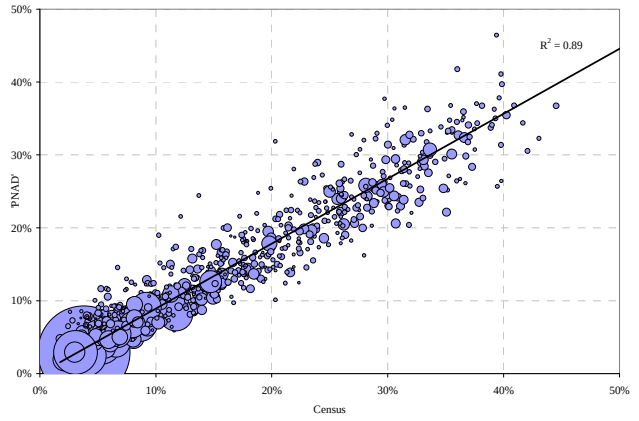

POF pseudo-sample



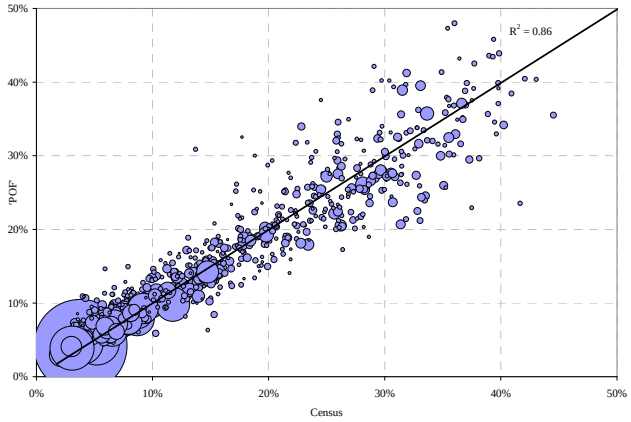

Source: Authors' Calculation.

Figure 13 - FGT(1) measures at Municipality level, ELL simulations and True Census estimates

PNAD pseudo-sample



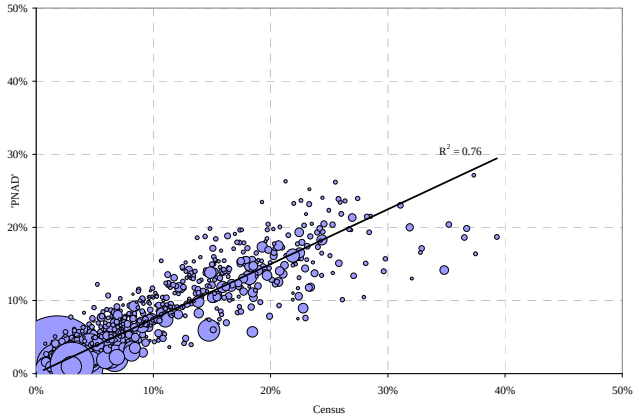

POF pseudo-sample



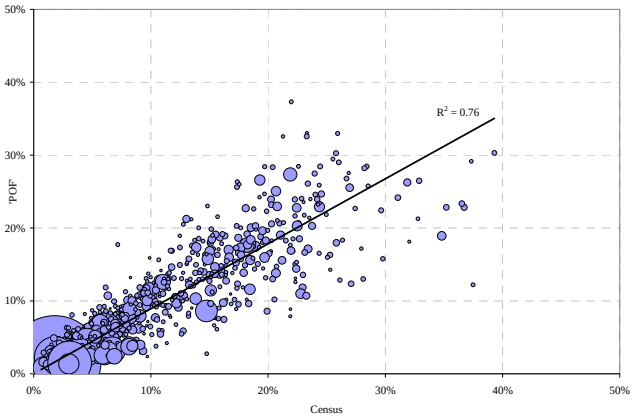

Source: Authors' Calculation.

Figure 14 - FGT(2) measures at Municipality level, ELL simulations and True Census estimates

PNAD pseudo-sample




POF pseudo-sample




Source: Authors' Calculation.

Given that standard errors accompany poverty estimates in the ELL methodology, a confidence interval can be drawn around each municipality level estimate in each pseudo-survey simulation. Tarozzi and Deaton (2007) focus their attention on this aspect of the ELL methodology – arguing that standard errors are likely to be too small. The smaller the estimated standard errors, the more narrow the confidence interval around each municipality-level poverty estimate. We thus ask whether, for each one of simulated welfare measures, we can verify whether the 95% confidence interval generated encompasses the 'true' welfare measure.

There are two ways to proceed with this coverage test as follows:

1. For each one of the 41 poverty maps constructed at the municipality level, we count how many municipalities out of 853 in the state have simulated confidence intervals that contain

the ‘true’ value of welfare measure obtained from the census. We present figures that indicate the share of municipalities containing the true value out of 853 municipalities for each survey;

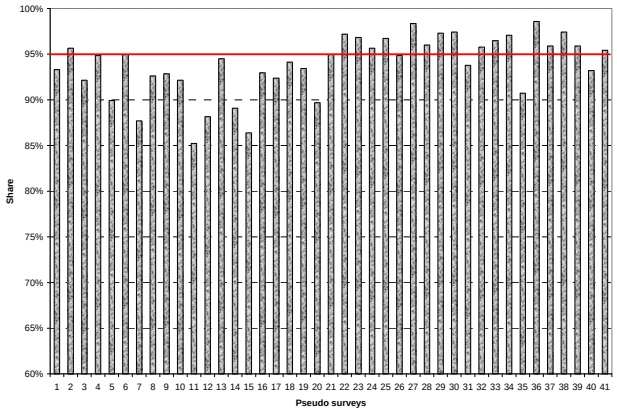
2. For each municipality in the state, we compute the fraction of simulations, 41 poverty maps, for which the simulated confidence intervals encompasses the ‘true’ value of the welfare measure. We present a histogram of the share of good predictions among the 853 municipalities of the state.

For both cases, we have just under 35,000 municipality-level poverty estimates deriving from our 41 poverty maps and 853 municipalities (853*41). Organizing these cells into a matrix form we have a matrix of 853 rows and 41 columns containing values one or zero depending on whether or not the municipality confidence interval contains the true welfare indicator. Let $K_{ij}$ represent these indicators, where *i=1,…,853* and *j=1,…,41*. Case 1 is obtained by the following average: $\overline{K}_{.j} = \sum_{i=1}^{853} K_{ij} \Big/ 853$, representing the sum of $K_{ij}$ over *i* for survey *j*; Case 2 is obtained in two steps: first we compute the average over surveys *j* $\overline{K}_{i.} = \sum_{j=1}^{41} K_{ij} \Big/ 41$ and then record its distribution.

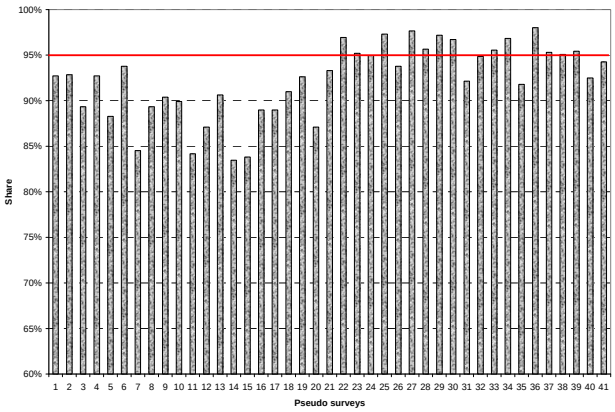
Figure 15 indicates that, for case 1, our predictions are quite robust and in general our confidence interval at the municipality level encompass the true point estimate in more than 90% of the cases for the FGT(0) and FGT(1) but nearer 80% for the FGT(2) in all surveys. The relatively low success in the case of the FGT(2) measure appears in part related to the fact that at the poverty line we are using (R$100 per person per month), FGT(2) values are very low in absolute terms, and this affects also calculations of precision. When we recalculate poverty based on a poverty line three times higher, coverage rates for the FGT(2) measure average around 90% rather than 80%.

Figure 15 - Share of municipalities where 95% confidence interval of ELL simulations encompass the actual FGT(α) measure for each pseudo-sample

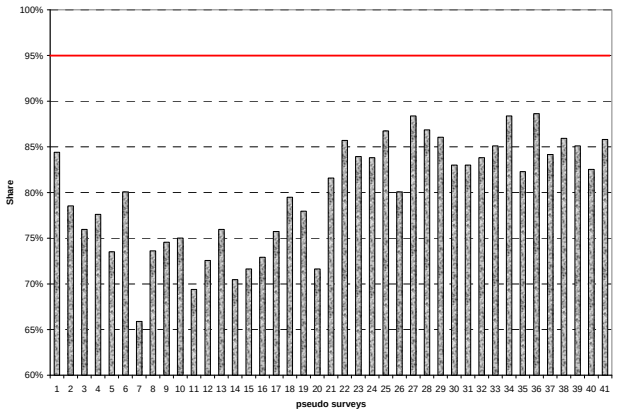
FGT(0)




FGT(1)



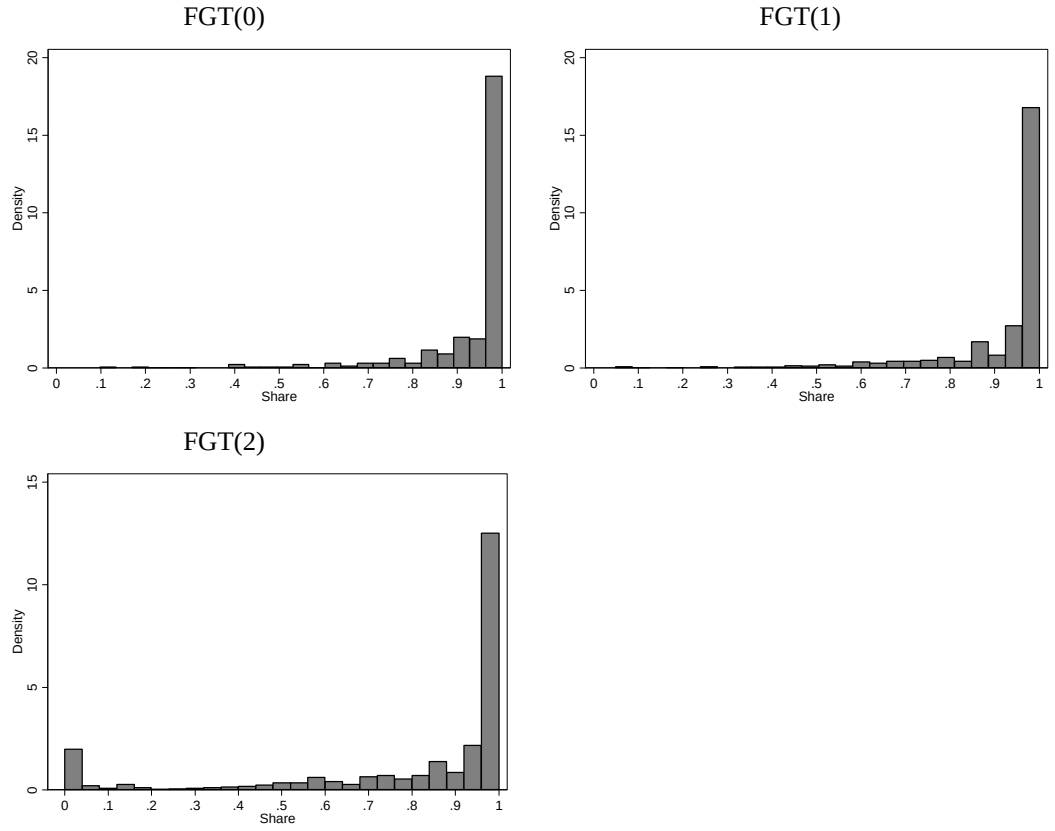

FGT(2)




Source: Authors' Calculation.

For the second case, figure 16 indicates that for the vast majority of municipalities a confidence interval of 95% or higher is sufficient to ensure that the "truth" is included in the confidence interval around the municipality level estimate. Hence, the claim that the 95% confidence intervals around our poverty map estimates will include the 'truth' 95 out of a 100 times does not appear unreasonable – particularly for FGT(0) and FGT(1) measures.

Figure 16: Histogram of the number of good predictions of ELL method at municipality level

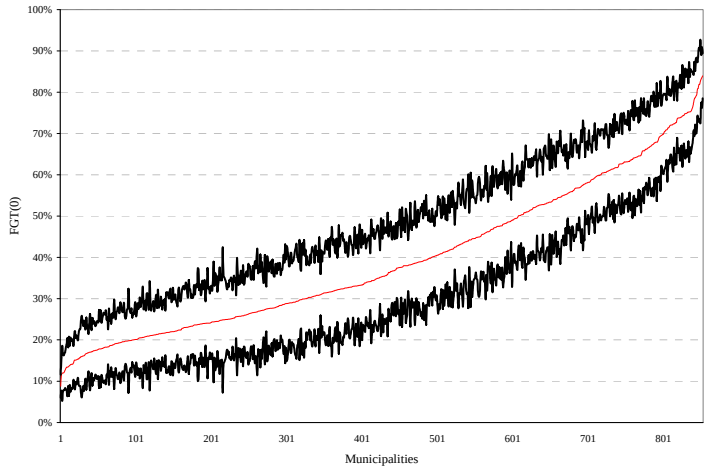



Source: Authors' Calculation.

Despite the encouraging results above, one might worry about the size of the confidence intervals we are simulating using the ELL approach – particularly as the conservative option of applying the cluster-effect at the municipality level in the simulations has been adopted. To what extent are the estimated standard errors sufficiently small to permit meaningful comparisons of poverty across municipalities? Coverage rates could be very high, as seen above, but if this is due to the standard errors being very large then the usefulness of the poverty map estimates becomes less obvious. Tarozzi and Deaton (2007) point out that, at least in principle, standard errors from the ELL method can explode if a sufficiently large intra-cluster correlation effect is applied entirely at the target population level. Figure 17 illustrates how estimated poverty varies across municipalities in the state based on estimates from one arbitrarily selected pseudo-survey and indicates that while, indeed, confidence errors are sufficiently large as to prevent fine pairwise comparisons of poverty across municipalities, there is a non-negligible number of municipalities

that are clearly distinguishable from one another in statistical terms. Figure 18 shows that at a conventional significance rate of 95% approximately 35% of municipalities can be ranked and distinguished in a statistical sense from one another. At a less stringent significance level of 75%, the number of significant rankings increases to 43%.[15]

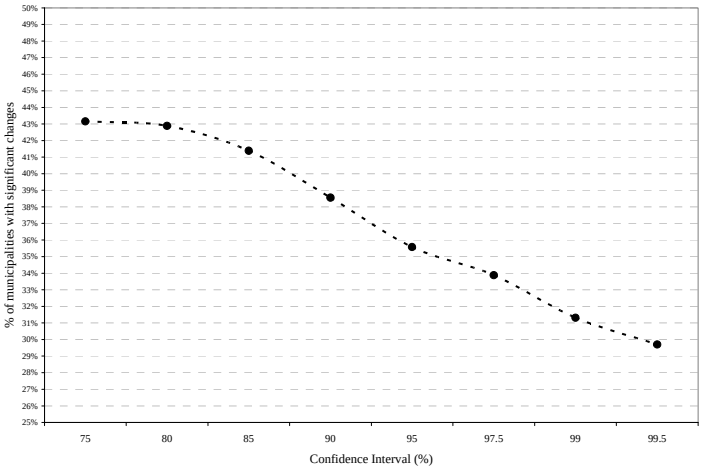
Figure 17: ELL simulation at municipality level and 95% confidence intervals




Source: Authors' Calculation.

[15] Does 35% represent an unacceptably low percentage of statistically significant rankings? It is not obvious that if municipality level, representative, household survey data were available, the proportion of statistically significant rankings would be much higher. For many pairwise comparisons of poverty, estimated poverty rates are very close and would require extremely precise estimates in order to yield statistically significant rankings. To illustrate, our 12.5% sample of the Census generates poverty measures at municipality level that yield statistically significant rankings for only 37% of all municipalities (considering, here, that the sample Census is associated with sampling error). Back of the envelope calculations that impose the 'typical' precision of stratum-level poverty estimates in Brazil's PNAD surveys on our municipality level poverty map estimates, indicate that the proportion of statistically significant rankings of municipalities (at 95% confidence levels) would be even less than the 35% reported here.

Figure 18: Share of municipalities that can be ranked and statistically distinguished one another



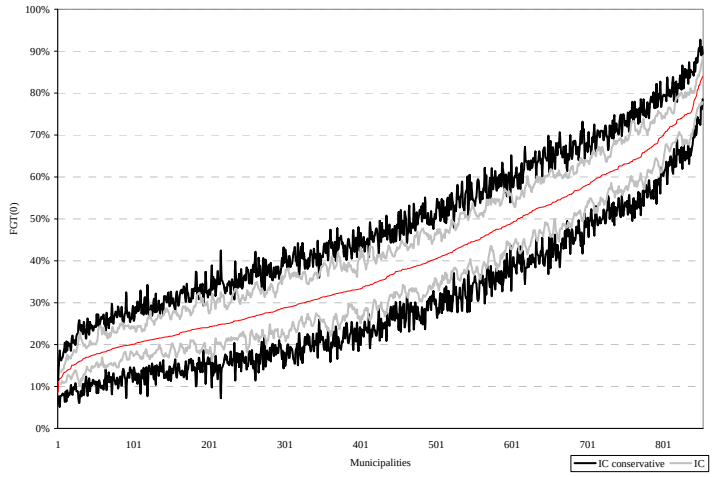

Source: Authors' Calculation.

Tarozzi and Deaton (2007) direct their concerns to a large extent at what they term the "standard" application of the ELL approach (what we call the "optimistic" approach, above) which in the context of this study would involve applying the location effect at the EA level in the Census rather than the municipality level. In the presence of some unobserved inter-cluster correlation, it is clear that standard errors obtained under the assumption of no such correlation will be understated. ELL (2002) found that, in the case of Ecuador, the degree of understatement was negligible if the inter-cluster correlation was assumed not to extend all the way to the target population level. Here, we have allowed for inter-cluster correlation all the way to the municipality level, and have indeed imposed the highly conservative assumption that the entire inter household correlation observed at the EA level applies at the municipality level (ELL conservative approach)

It is of interest to ask how far wrong we might have gone in our Minas Gerais setting if we had not taken this conservative stance, and had proceeded with the "optimistic" approach of applying the location effect entirely at the EA level. Figure 19 indicates that, indeed, confidence intervals are narrower when the location effects are applied only at the EA level. However, it is noteworthy that a hypothetical policy maker, presented with such an "optimistic" poverty map and its accompanying standard errors, would not come away with a wildly unreasonable picture of the

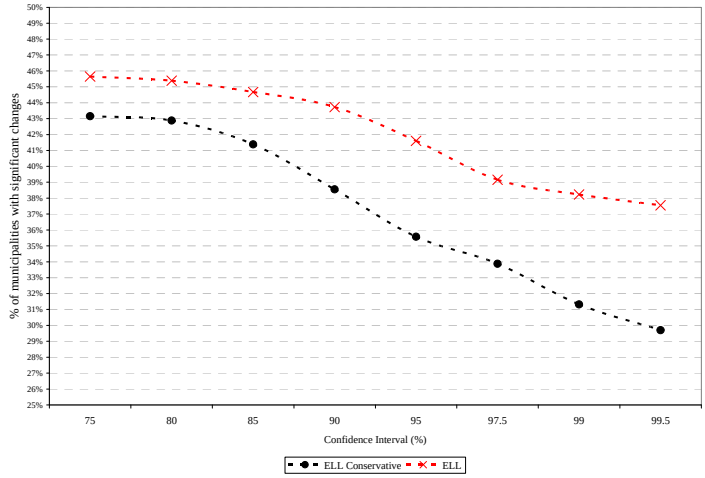
spatial distribution of poverty. Figure 20 illustrates that instead of observing that 35% of municipalities can be ranked (with 95% confidence), the "optimistic" poverty map would have led the policy maker to perceive 42% of municipalities as rankable. At a 75% confidence level the difference between the "optimistic" and "conservative" poverty maps is less than three percentage points.

Figure 19: ELL simulation at municipality level and 95% confidence intervals




Source: Authors' Calculation.

Figure 20: Share of municipalities that can be ranked and statistically distinguished one another




Source: Authors' Calculation.

## VI. Conclusion

The results presented here suggest that, in a setting where the underlying assumptions can be explicitly scrutinized and estimates can be compared to their "true" values, the ELL methodology performs reasonably well. In the state of Minas Gerais, Brazil, a unique model estimated at the

state level with sample survey data yields parameter estimates that are used to impute incomes to individual households in the population census. These imputed incomes can then be examined to assess poverty at the level of each of the nearly 1000 municipalities in the state. The resultant poverty estimates have been found to line up quite closely to the actual, observed, poverty rates in those municipalities. Moreover, confidence intervals around the point estimates are both moderate in size and encompass the "truth" broadly in accordance with the statistical precision they are intended to reflect. A recent critique of the ELL methodology by Tarozzi and Deaton (2007) uses Monte Carlo simulations to illustrate that there are conditions and circumstances under which the ELL methodology can yield a sense of precision that is far too optimistic. The present study has shown that in an empirical setting that one might have thought, a priori, would fit the conditions for the Tarozzi and Deaton (2007) critique, the ELL methodology performs quite well. The broad applicability of the Tarozzi and Deaton (2007) argument is thus brought into question. Further empirical research into these questions is needed, in different settings and circumstances.

**References**


Araujo, C., Ferreira, F. H. G., Lanjouw, P. And Ozler, B. (forthcoming) 'Local Inequality and Project Choice: Theory and Evidence from Ecuador' mimeo, the World Bank, forthcoming *Journal of Public Economics*.

Alderman, H., Babita, M., Demombynes, G., Makhatha, N., Ozler, B. (2002) "How Low Can You Go? Combining Census and Survey Data for Mapping Poverty in South Africa" *Journal of African Economies*, 11(2), 169-200.

Christiaensen, L. and Stifel, D. (2007) 'Tracking Poverty Over Time in the Absence of Comparable Consumption Data', *World Bank Economic Review*, 21(2):317-341.

Demombynes, G., Lanjouw, J.O., Lanjouw, P. and Elbers (2006) "How Good a Map? Putting Small Area Estimation to the Test", Policy Research Working Paper No. 4155, The World Bank.

Demombynes, G. and Ozler, B. (2005) 'Crime and Local Inequality in South Africa', *Journal of Development Economics*, 76(2): 265-292.

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

Elbers, C., Lanjouw, J.O., Lanjouw, P. (2002) "Micro-Level Estimation of Welfare" Policy Research Working Paper 2911, DECRG, The World Bank.

Elbers, C., Fujii, T., Lanjouw, P., Ozler, B. and Yin, W. (2007) 'Poverty Alleviation Through Geographic Targeting: How Much Does Disaggregation Help?' *Journal of Development Economics*, 83(1): 198-213.

Fujii, T. (2005) 'Micro-Level Estimation of Child Malnutrition Indicators and its Application to Cambodia' Policy Research Working Paper No. 3552, The World Bank.

Fujii, T. and Roland-Holst, D. (2007) 'How Does Vietnam's Accession to the World Trade Organization Change the Spatial Incidence of Poverty?' UNU-Wider Research Paper No. 2007/12, United Nations University – World Insititute for Development Economics Research, Helsinki, Finland.

Kijima, Y., and Lanjouw, P. (2003) 'Poverty in India During the 1990s: A Regional Perspective', Policy Research Working Paper No. 3141, the World Bank.

Pfefferman and Tiller (2005) 'Bootstrap Approximation to Prediction MSE for State-Space Models with Estimated Parameters' *Journal of Time Series Analysis* , 26: 893-216.

Tarozzi, A. and Deaton, A. (2007) 'Using Census and Survey Data to Estimate Poverty and Inequality in Small Areas', mimeo, Duke University, North Carolina.

**Appendix 1: Spatial correlation of the error**

An alternative perspective on the analysis in Section V.3 can be obtained if we directly analyze the spatial correlation of the unexplained component of simulated income on the basis of the Moran's I spatial correlation statistic. Moran's I statistic tests for global spatial autocorrelation - the degree to which nearby areas have similar error terms. Where these are similar in nearby areas, Moran's I will be large and positive. In contrast, dissimilar rates generate a negative Moran's I statistic. Mathematically, the Local Moran's I statistic is represented by the spatial regression of the error component $\overline{Lny_c} - \overline{\hat{Lny}_c}$ on the spatial lagged error component $W \cdot \left[\overline{Lny_c} - \overline{\hat{Lny}_c}\right]$.[16] We provide a graphical presentation to highlight spatial autocorrelation for each individual geographical location. We estimate the model at the state-level using an arbitrarily selected pseudo-sample. We first apply a specification that does not control for local characteristics. We then re-estimate the model with enumeration area aggregates included as regressors. We then compute $\overline{\hat{\log y_c}} = \overline{\mathbf{x}'_{CENSUS_c}} \cdot \tilde{\tilde{\beta}}^{r}_{SURVEY}$ and finally we compute the error components.

On the basis of the census data we estimate the model at state level R but we only present results here for the capital Belo Horizonte because of its size and heterogeneity and also because a state map at the enumeration area level would be very difficult to read. Figures A1.1 and A1.2 illustrate how heterogeneous enumeration areas are in terms of income in Belo Horizonte, and also their spatial correlation (based on Rook Contiguity). The southern enumeration areas in the municipality are rich and their positive correlation is represented in a high-high (red) significant estimator. In contrast the north is also positive and significantly correlated but these are poorer than in the south. Here we have a low-low correlation. The estimated Moran's I statistic is significant and equal to 0.4481.

[16] Moran's I statistic is the average of Local Moran's I statistics.

Figure A1.1: Spatial distribution of the log $y_c$

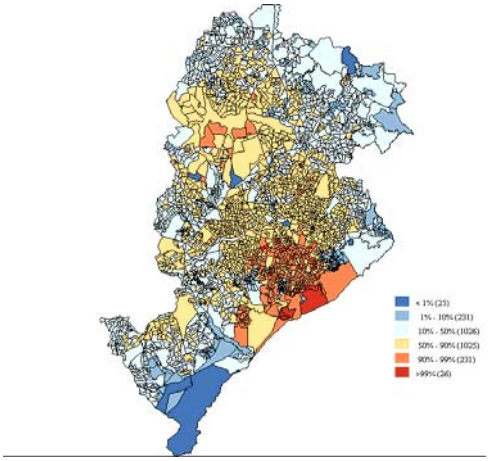
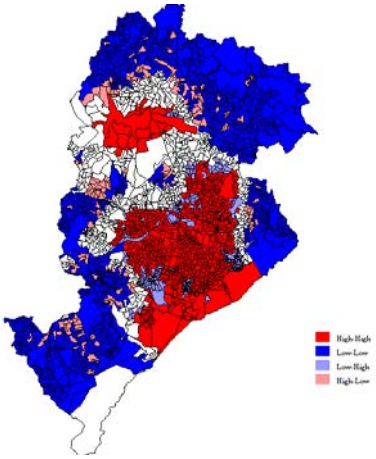



Figure A1.2 - Local Moran's I statistic (LISA) of the log $y_c$



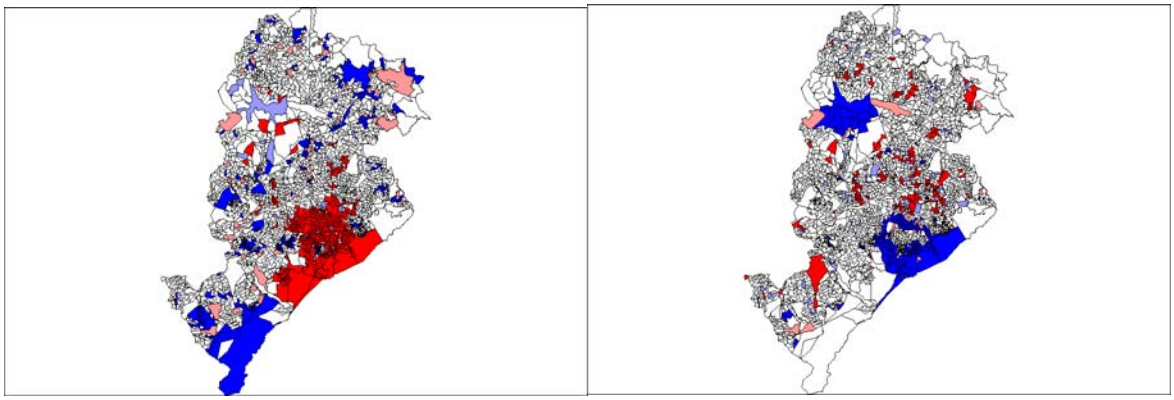

Source: Authors' calculation.

The figures in A1.3 below, referring to Model A, illustrate local spatial correlation when the estimated model does not control for location characteristics. Model B refers to the case where EA-level aggregates have been added. Local spatial correlation of the error component is pronounced across 6 different weighting schemes when we perform the naïve model (Model A). The addition of local controls demonstrates that the high-high (red) and low-low (dark blue) became much less present in any of the weighting schemes.

Figure A1.3 – Local Moran's statistics of the error component obtained by a state level model A (**WITHOUT** any enumeration area aggregates) and model B (**WITH** enumeration area aggregates)

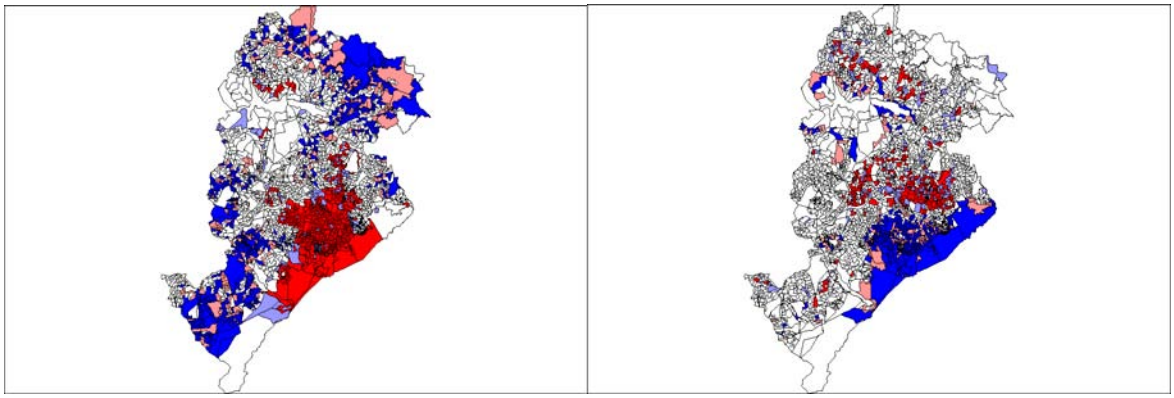
Rook Contiguity: order 1

Model A | Model B



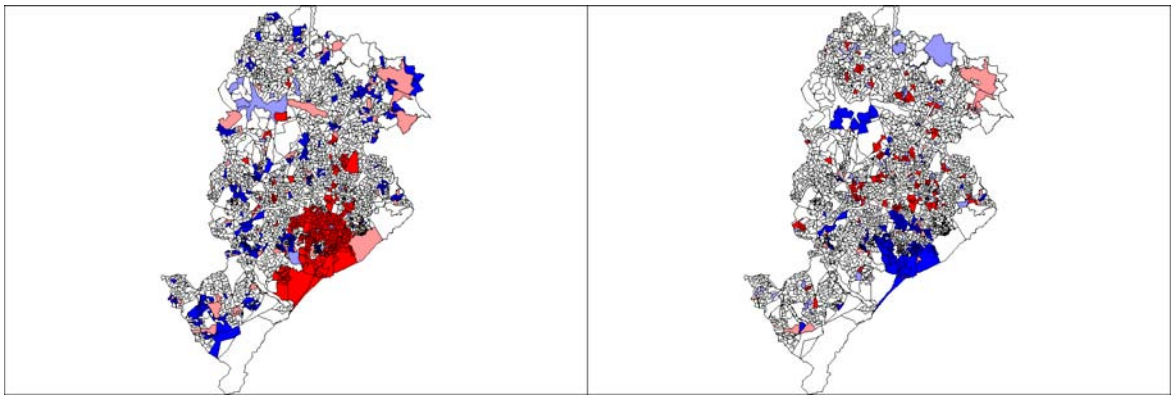
Rook Contiguity: order 3

Model A | Model B



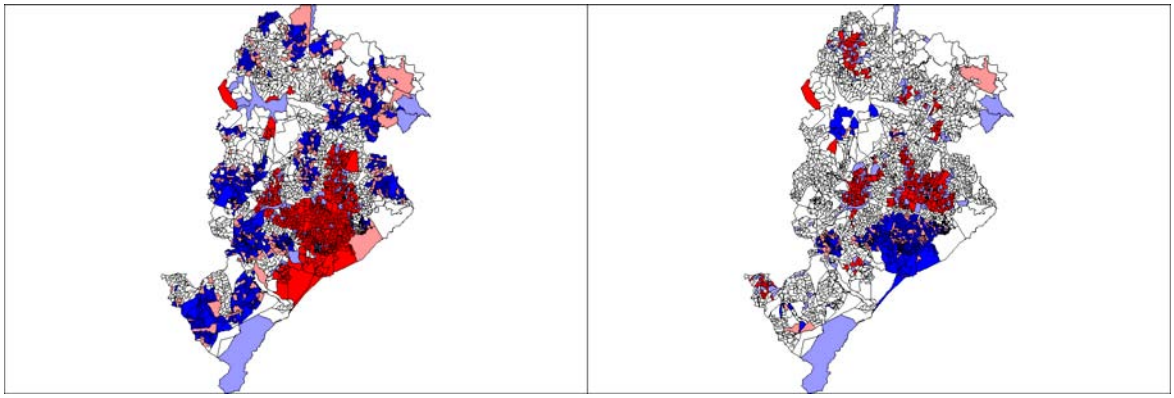
5 Nearest neighbor

Model A | Model B



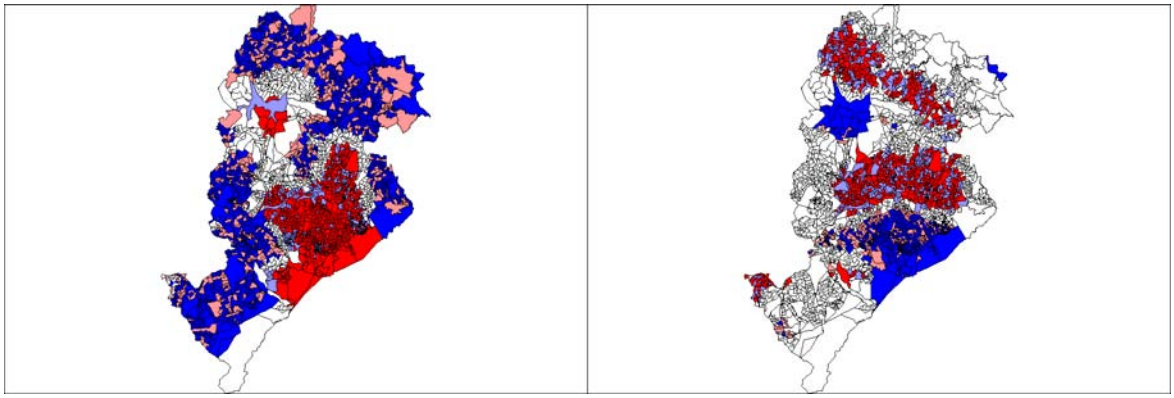
Euclidian distance 1km

Model A | Model B



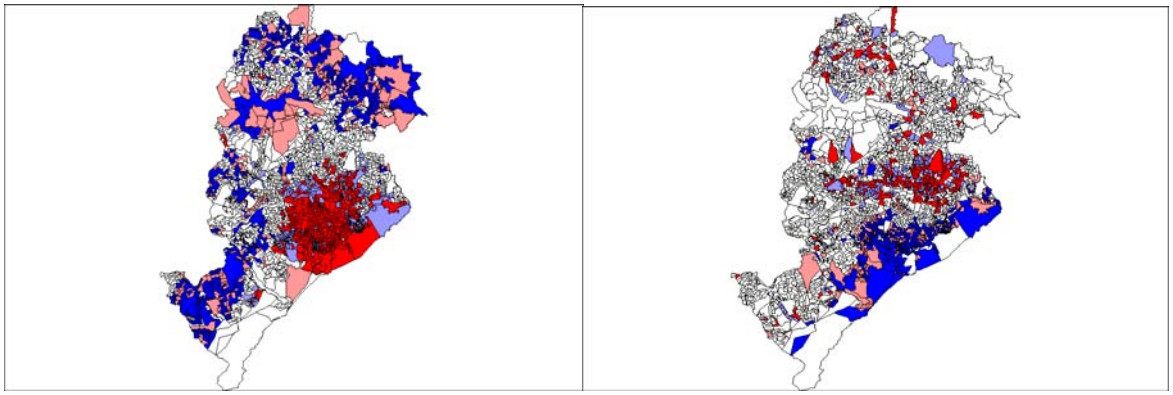
Euclidian distance 2.1km

Model A | Model B



Queen Contiguity order 5

Model A | Model B



Source: Authors' calculation.

## Appendix 2:
## First-Stage Regression Models: PNAD and POF Examples

| **Log Per Capita Income Model: PNAD-type sample:** | **Coefficient** | **Std. Err.** | **t** | **\|Prob\|>t** |
|---|---|---|---|---|
| Constant | 5.904 | 0.147 | 40.055 | 0.000 |
| Household level variables | | | | |
| water service: regular | 0.061 | 0.026 | 2.350 | 0.019 |
| 0 year of schooling | -0.359 | 0.038 | -9.528 | 0.000 |
| 1-3 year of schooling | -0.200 | 0.019 | -10.448 | 0.000 |
| 8-10 year of schooling | 0.287 | 0.023 | 12.563 | 0.000 |
| 11-14 year of schooling | 0.530 | 0.023 | 22.919 | 0.000 |
| 14-+ year of schooling | 0.900 | 0.035 | 25.654 | 0.000 |
| Owner of household | -0.044 | 0.020 | -2.234 | 0.026 |
| Other status of ownership | -0.225 | 0.027 | -8.205 | 0.000 |
| don't have bathroom inside | -0.152 | 0.039 | -3.933 | 0.000 |
| do have washing machine | 0.254 | 0.017 | 14.804 | 0.000 |
| Paved street | 0.041 | 0.020 | 2.062 | 0.039 |
| Black | -0.117 | 0.014 | -8.208 | 0.000 |
| Not attending school but studied | -0.102 | 0.033 | -3.134 | 0.002 |
| don't have sewer public service | -0.090 | 0.020 | -4.425 | 0.000 |
| don't have refregirator | -0.306 | 0.022 | -13.906 | 0.000 |
| female | -0.190 | 0.026 | -7.211 | 0.000 |
| don't have microwave | -0.388 | 0.021 | -18.105 | 0.000 |
| age | 0.012 | 0.001 | 18.449 | 0.000 |
| 0 people older than 65 | 0.045 | 0.025 | 1.828 | 0.068 |
| 2 adults over 65 years-old | -0.138 | 0.037 | -3.726 | 0.000 |
| Type of household: house | -0.079 | 0.030 | -2.684 | 0.007 |
| Type of household: single room | -0.292 | 0.071 | -4.087 | 0.000 |
| Family type 1 | 0.491 | 0.035 | 13.918 | 0.000 |
| Family type 2 | 0.214 | 0.035 | 6.179 | 0.000 |
| Family type 3 | -0.277 | 0.030 | -9.195 | 0.000 |
| Family type 5 | -0.185 | 0.036 | -5.111 | 0.000 |
| Municipality level variables | | | | |
| Share of households with proper garbage collection | 0.386 | 0.066 | 5.824 | 0.000 |
| Share of migrants | -0.103 | 0.055 | -1.858 | 0.063 |
| District level variables | | | | |
| Average schooling of population | -0.139 | 0.020 | -6.927 | 0.000 |
| Share of people out of labour force | -0.425 | 0.160 | -2.653 | 0.008 |
| Share of self-employed | -0.268 | 0.143 | -1.878 | 0.060 |
| Average income per capita | 0.000 | 0.000 | 6.003 | 0.000 |
| E.A. Level variables | | | | |
| Average head's income | 0.000 | 0.000 | 6.568 | 0.000 |
| Average years of schooling of heads | 0.046 | 0.007 | 6.619 | 0.000 |
| Average of household size | -0.076 | 0.021 | -3.627 | 0.000 |

Number of Observations used in the Model=11704
Number of Records in the dataset=12122
MSE=0.5035
RMSE=0.7096
R2=0.5572

**Heterocesdasticity model: PNAD**

| | Coefficient | Std. Err. | t | \|Prob\|>t |
|---|---|---|---|---|
| Constant | -5.238 | 0.205 | -25.529 | 0.000 |
| 14-+ year of schooling | -0.258 | 0.103 | -2.503 | 0.012 |
| Iliterate | -0.162 | 0.068 | -2.378 | 0.017 |
| Share of informal employees*_yhat_ | -0.186 | 0.064 | -2.915 | 0.004 |
| Share of formal employees | -1.058 | 0.357 | -2.967 | 0.003 |
| Number of hospitals | 0.003 | 0.001 | 2.390 | 0.017 |
| Age*_yhat_*_yhat_ | 0.000 | 0.000 | 7.351 | 0.000 |
| zero members age above 65 years-old | 0.289 | 0.070 | 4.131 | 0.000 |
| Share of households with proper garbage collection | -0.522 | 0.189 | -2.760 | 0.006 |
| Type of household: single room | 0.372 | 0.207 | 1.793 | 0.073 |
| Variance of years of schooling of heads at EA level | 0.015 | 0.005 | 3.072 | 0.002 |

| **Log Income Per Capita: POF-type sample** | **Coefficient** | **Std. Err.** | **t** | **\|Prob\|>t** |
|---|---|---|---|---|
| Constant | 3.824 | 0.235 | 16.263 | 0.000 |
| Household level variables | | | | |
| 0 year of schooling | -0.184 | 0.047 | -3.953 | 0.000 |
| 4-7 year of schooling | 0.216 | 0.037 | 5.800 | 0.000 |
| 8-10 year of schooling | 0.526 | 0.053 | 9.941 | 0.000 |
| 11-14 year of schooling | 0.818 | 0.053 | 15.519 | 0.000 |
| 14-+ year of schooling | 1.418 | 0.073 | 19.494 | 0.000 |
| Capital city: NO | 0.240 | 0.069 | 3.505 | 0.001 |
| White | 0.095 | 0.029 | 3.300 | 0.001 |
| Attending school | 0.416 | 0.133 | 3.127 | 0.002 |
| do have refregirator | 0.261 | 0.044 | 5.895 | 0.000 |
| Age | 0.014 | 0.001 | 12.624 | 0.000 |
| do have microwave | 0.543 | 0.043 | 12.692 | 0.000 |
| Number of children | -0.165 | 0.020 | -8.131 | 0.000 |
| Paved street | 0.119 | 0.035 | 3.437 | 0.001 |
| metropolitan region: yes | 0.171 | 0.045 | 3.780 | 0.000 |
| Male | 0.162 | 0.055 | 2.952 | 0.003 |
| do have TV | 0.159 | 0.050 | 3.199 | 0.001 |
| Type of household: house | 0.550 | 0.142 | 3.883 | 0.000 |
| Type of household: apartment | 0.829 | 0.150 | 5.538 | 0.000 |
| Family type 1 | 0.182 | 0.066 | 2.773 | 0.006 |
| Family type 3 | -0.176 | 0.045 | -3.905 | 0.000 |
| Family type 5 | -0.173 | 0.073 | -2.364 | 0.018 |
| Family type 6 | -0.116 | 0.080 | -1.453 | 0.146 |
| household size | -0.076 | 0.010 | -7.703 | 0.000 |
| District level variables | | | | |
| Share of people out of labour force | -1.214 | 0.308 | -3.938 | 0.000 |
| Share of people unemployed | -0.847 | 0.260 | -3.262 | 0.001 |
| Share of informal employees | 0.682 | 0.176 | 3.876 | 0.000 |
| Average income per capita | 0.000 | 0.000 | 1.654 | 0.098 |

Number of Observations used in the Model=2790
Number of Records in the dataset=2880
MSE=0.4892
RMSE=0.6995
R2=0.5814

**Heteroscedasticity model: POF**

| | Coefficient | Std. Err. | t | \|Prob\|>t |
|---|---|---|---|---|
| _intercept_ | -4.7631 | 0.2109 | -22.5827 | 0 |
| 8-10 year of schooling | 0.4679 | 0.1443 | 3.2429 | 0.0012 |
| White | 0.1459 | 0.1889 | 0.7724 | 0.4399 |
| White*_yhat_*_yhat_ | 0.0046 | 0.0033 | 1.4081 | 0.1592 |
| Share of informal employees | -1.8613 | 0.4964 | -3.7499 | 0.0002 |
| Do not have refrigerator*_yhat_*_yhat_ | 0.0184 | 0.0047 | 3.896 | 0.0001 |
| Average income per capita of the district | 0.0002 | 0.0001 | 1.771 | 0.0767 |
| Female | 0.3217 | 0.1265 | 2.5435 | 0.011 |
| Proper sanitation | -0.1487 | 0.114 | -1.3035 | 0.1925 |
| Family type 1 | -0.2015 | 0.2072 | -0.9725 | 0.3309 |
| Family type 1*_yhat_*_yhat_ | -0.0033 | 0.0016 | -2.0616 | 0.0393 |