POLICY RESEARCH WORKING PAPER 2911

Micro-Level Estimation of Welfare

Chris Elbers
Jean O. Lanjouw
Peter Lanjouw

The World Bank
Development Research Group
Poverty Team
October 2002

POLICY RESEARCH WORKING PAPER 2911

Abstract

The authors construct and derive the properties of estimators of welfare that take advantage of the detailed information about living standards available in small household surveys and the comprehensive coverage of a census or large sample. By combining the strengths of each, the estimators can be used at a remarkably disaggregated level. They have a clear interpretation, are mutually comparable, and can be assessed for reliability using standard statistical theory.

Using data from Ecuador, the authors obtain estimates of welfare measures, some of which are quite reliable for populations as small as 15,000 households—a "town." They provide simple illustrations of their use. Such estimates open up the possibility of testing, at a more convincing intra-country level, the many recent models relating welfare distributions to growth and a variety of socioeconomic and political outcomes.

This paper—a product of the Poverty Team, Development Research Group—is part of a larger effort in the group to develop tools for the analysis of poverty and income distribution. Copies of the paper are available free from the World Bank, 1818 H Street NW, Washington, DC 20433. Please contact Patricia Sader, room MC3-556, telephone 202-473-3902, fax 202-522-1153, email address psader@worldbank.org. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at celbers@econ.vu.nl, jlanjouw@brookings.edu, or planjouw@worldbank.org. October 2002. (57 pages)

MICRO-LEVEL ESTIMATION OF WELFARE

BY CHRIS ELBERS, JEAN O. LANJOUW, AND PETER LANJOUW[1]

[1]We are very grateful to Ecuador's Instituto Nacional de Estadistica y Censo (INEC) for making its 1990 unit-record census data available to us. Much of this research was done while the authors were at the Vrije Universiteit, Amsterdam, and we appreciate the hospitality and input from colleagues there. We also thank Don Andrews, François Bourguignon, Andrew Chesher, Denis Cogneau, Angus Deaton, Jean-Yves Duclos, Francisco Ferreira, Jesko Hentschel, Michiel Keyzer, Steven Ludlow, Berk Özler, Giovanna Prennushi, Martin Ravallion, Piet Rietveld, John Rust and Chris Udry for comments and useful discussions, as well as seminar participants at the Vrije Universiteit, ENRA (Paris), U.C. Berkeley, Georgetown University, the World Bank and the Brookings Institution. Financial support was received from the Bank Netherlands Partnership Program.

1. INTRODUCTION

RECENT THEORETICAL ADVANCES have brought income and wealth distributions back into a prominent position in growth and development theories, and as determinants of specific socio-economic outcomes, such as health or levels of violence.[2] Empirical investigation of the importance of these relationships, however, has been held back by the lack of sufficiently detailed high quality data on distributions. Time series data are sparse, constraining most econometric analyses to a cross-section of countries. Not only may these data be non-comparable, such estimations require strong assumptions about the stability of structural relationships across large geographical areas and political units.[3] Further, many of the hypothesized relationships are more obviously relevant for smaller groups or areas. For example, as noted by Deaton (1999), while it is not clear why country-wide

[2]The models in this growing literature describe a wide variety of linkages between distributions and growth. For example, inequality (or poverty) limits the size of markets which slows growth when there are scale economies (Murphy, Shleifer and Vishny, 1989); with imperfect capital markets, greater inequality limits those able to make productive investment and occupational choices (Galor and Zeira, 1993; Banerjee and Newman, 1993). Aghion and Bolton (1997) endogenize inequality, with growth having a feedback effect on the distribution of wealth via its effect on credit, or labour, markets. Political economy models such as Alesina and Rodrik (1994) and Persson and Tabellini (1994) suggest that, in democratic regimes, inequality will lead to distortionary redistributive policies which slow growth.

[3]The state-of-the-art data set for this purpose, compiled by Deininger and Squire (1996), goes a long way towards establishing comparability but the critique by Atkinson and Brandolini (2001) shows it remains very far from ideal. (See also Fields, 1989 and 2001, on data.)

Bruno, Ravallion and Squire (1998) give examples of country-level estimation of growth models. Although they do not include distributional variables, Barro and Sala-i-Martin estimate a growth model using U.S. state-level data where the fact that it is a better controlled situation is emphasized (see Comments and Discussion in Barro and Sala-i-Martin, 1991). Ravallion (1998) points out that aggregation alone can bias estimates of the relationship between asset inequality and income growth derived from country-level data, and demonstrates this using *county*-level panel data from China. For a more general identification critique of cross-country models see Banerjee and Duflo (2000).

inequality should directly affect an individual's health, a link could be made to the degree of inequality within his reference group.

The problem confronted is that household surveys that include reasonable measures of income or consumption can be used to calculate distributional measures, but at low levels of aggregation these samples are rarely representative or of sufficient size to yield statistically reliable estimates. At the same time, census (or other large sample) data of sufficient size to allow disaggregation either have no information about income or consumption, or measure these variables poorly.[4] This paper outlines a statistical procedure to combine these types of data to take advantage of the detail in household sample surveys and the comprehensive coverage of a census. It extends the literature on small area statistics (Ghosh and Rao (1994), Rao (1999)) by developing estimators of population parameters which are non-linear functions of the underlying variable of interest (here unit level consumption), and by deriving them from the full unit level distribution of that variable.

In examples using Ecuadorian data, our estimates have levels of precision comparable to those of commonly used survey based welfare estimates - but for populations as small as 15,000 households, a 'town'. This is an enormous improvement over survey

[4]For example, a single question regarding individuals' incomes in the 1996 South African census generates an estimate of national income just 83% the size of the national *expenditure* estimate derived from a representative household survey, and a per-capita poverty rate 25% higher, with discrepancies systematically related to characteristics such as household location (Alderman, *et.al.*, 2002).

based estimates, which are typically only consistent for areas encompassing hundreds of thousands, even millions, of households. Experience using the method in South Africa, Brazil, Panama, Madagascar and Nicaragua suggest that Ecuador is not an unusual case (Alderman, *et. al.* (2002), and Elbers, Lanjouw, Lanjouw, and Leite (2002)).

With accurate welfare measures for groups the size of towns, villages or even neighborhoods, researchers should be able to test hypotheses at an appropriate level of disaggregation, where assumptions about a stable underlying structure are more tenable. Better local measures of poverty and inequality will also be useful in the targetting of development assistance and many governments are enthusiastic about new methods for using their survey and census data for this purpose. Poverty 'maps' can be simple and effective policy tools. Disaggregated welfare estimates can also help governments understand the tradeoffs involved in decentralizing their spending decisions. While it is beneficial to take advantage of local information about community needs and priorities, if local inequalities are large and decisions are taken by the elite, projects may not benefit the poorest. Local level inequality measures, together with data on project choices, make it possible to shed light on this potential cost of decentralization.

Datasets have been combined to fill in missing information or avoid sampling biases in a variety of other contexts. Examples in the econometric literature include Arellano and Meghir (1992) who estimate a labour supply model combining two samples. They

use the UK Family Expenditure Survey (FES) to estimate models of wages and other income conditioning on variables common across the two samples. Hours and job search information from the much larger Labour Force Survey is then supplemented by predicted financial information. In a similar spirit, Angrist and Krueger (1992) combine data from two U.S. censuses. They estimate a model of educational attainment as a function of school entry age, where the first variable is available in only in one census and the second in another, but an instrument, birth quarter, is common to both. Lusardi (1996) applies this two-sample IV estimator in a model of consumption behaviour. Hellerstein and Imbens (1999) estimate weighted wage regressions using the U.S. National Longitudinal Survey, but incorporate aggregate information from the U.S. census by constructing weights which force moments in the weighted sample to match those in the census.

After the basic idea is outlined, we develop a model of consumption in Section 3. We use a flexible specification of the disturbance term that allows for non-normality, spatial autocorrelation and heteroscedasticity. One might ask whether, given a reasonable first-stage model of consumption, it would suffice to calculate welfare measures on the basis of predicted consumption alone. In general such an approach yields inconsistent estimates and, more importantly, it may not even preserve welfare rankings of villages. Figures 1.a and 1.b demonstrate using the data from Ecuador described below. In Figure 1.a 'villages' are ordered along the x-axis according to a consistent estimate of the expected

proportion of their households that are poor. The jagged line represents estimates of the same proportions based only on the systematic part of households' consumption. Figure 1.b shows the same comparison for the expected general entropy (0.5) measure of inequality. There is clearly significant and sizable bias and re-ranking associated with ignoring the unobserved component of consumption even with the extensive set of regressors available to us in this example. Thus one would expect the use of predicted consumption to be problematic in many actual applications.

The welfare estimator is developed in Section 4 and its properties derived in Section 5. Section 6 gives computational details with results for our Ecuadorian example presented in Section 7. In this section, we explore briefly the implications of making various modelling assumptions. Section 8 indicates how much the estimator improves on sample based estimates. Section 9 gives results for additional welfare measures and then, in Section 10, we provide simple illustrations of the use of our estimators. The final section concludes.

2. The Basic Idea

The idea is straightforward. Let W be an indicator of poverty or inequality based on the distribution of a household-level variable of interest, y_h. Using the smaller and richer data sample, we estimate the joint distribution of y_h and a vector of covariates, x_h. By restricting the set of explanatory variables to those that can also be linked to

households in the larger sample or census, this estimated distribution can be used to generate the distribution of y_h for any sub-population in the larger sample conditional on the sub-population's observed characteristics.[5] This, in turn, allows us to generate the conditional distribution of W, in particular, its point estimate and prediction error.

3. The Consumption Model

The first concern is to develop an accurate empirical model of y_{ch}, the per capita expenditure of household h in sample cluster c. We consider a linear approximation to the conditional distribution of y_{ch},

$$(1) \quad \ln y_{ch} = E[\ln y_{ch}|x_{ch}^T] + u_{ch} = x_{ch}^T\beta + u_{ch},$$

where the vector of disturbances $u \sim \mathcal{F}(0, \Sigma)$.[6] Note that, unlike in much of econometrics, β is not intended to capture only the direct effect of x on y. Because the survey estimates will be used to impute into the census, if there is (unmodelled) variation in the parameters we would prefer to fit most closely the clusters that represent large census populations. This argues for weighting observations by population expansion factors.

[5]The explanatory variables are observed values and thus need to have the same degree of accuracy in addition to the same definitions across data sources. Comparing distributions of reponses at a level where the survey is representative is a check that we have found to be important in practice.

[6]One could consider estimating E($y|x$) or the conditional density p($y|x$) non-parametrically. In estimating expenditure for each household in the populations of interest (perhaps totalling millions) conditioning on, say, thirty observed characteristics, a major difficulty is to find a method of weighting that lowers the computational burden. See Keyzer (2000) and Tarozzi (2002) for examples and discussion

To allow for a within cluster correlation in disturbances, we use the following specification:

$$u_{ch} = \eta_c + \varepsilon_{ch},$$

where η and ε are independent of each other and uncorrelated with observables, x_{ch}. One expects location to be related to household income and consumption, and it is certainly plausible that some of the effect of location might remain unexplained even with a rich set of regressors. For any given disturbance variance, σ^2_{ch}, the greater the fraction due the common component η_c the less one enjoys the benefits of aggregating over more households within a village. Welfare estimates become less precise. Furthur, the greater the part of the disturbance which is common, the lower will be inequality. Thus, failing to take account of spatial correlation in the disturbances would result in underestimated standard errors on welfare estimates, and upward biased estimates of inequality (but see the examples below).

Since residual location effects can greatly reduce the precision of welfare estimates, it is important to explain the variation in consumption due to location as far as possible with the choice and construction of x_{ch} variables. We see in the example below that location means of household-level variables are particularly useful. Clusters in survey data typically correspond to enumeration areas (EA) in the population census. Thus, means can be calculated over all households in an EA and merged into the smaller sample data.

Because they include far more households, location means calculated in this way give a considerably less noisy indicator than the same means taken over only the households in a survey cluster. Other sources of information could be merged with both census and survey datasets to explain location effects as needed. Geographic information system databases, for example, allow a multitude of environmental and community characteristics to be geographically defined both comprehensively and with great precision.

An initial estimate of β in equation (1) is obtained from OLS or weighted least squares estimation. Denote the residuals of this regression as $\widehat{u}_{ch}$. The number of clusters in a household survey is generally too small to allow for heteroscedasticity in the cluster component of the disturbance. However, the variance of the idiosyncratic part of the disturbance, $\sigma^2_{\varepsilon,ch}$, can be given a flexible form. With consistent estimates of β, the residuals e_{ch} from the decomposition

$$\widehat{u}_{ch} = \widehat{u}_{c.} + (\widehat{u}_{ch} - \widehat{u}_{c.}) = \widehat{\eta}_c + e_{ch},$$

(where a subscript '.' indicates an average over that index) can be used to estimate the variance of ε_{ch}. We propose a logistic form,

$$(2) \quad \sigma^2(z_{ch}, \alpha, A, B) = \left[\frac{Ae^{z_{ch}^T\alpha} + B}{1 + e^{z_{ch}^T\alpha}}\right]$$

The upper and lower bounds, A and B, can be estimated along with the parameter vector α using a standard pseudo maximum likelihood procedure.[7] This functional form avoids

[7] An estimate of the variance of the estimators can be derived from the information matrix and used to

both negative and extremely high predicted variances.

The variance, σ_η^2, of the remaining (weighted) cluster random effect is estimated non-parametrically, allowing for heteroscedasticity in ε_{ch}. This is a straightforward application of random effect modelling (e.g., Greene (2000), Section 14.4.2). An alternative approach based on moment conditions gives similar results. See Appendix 1.

In what follows we need to simulate the residual terms η and ε. Appropriate distributional forms can be determined from the cluster residuals $\widehat{\eta}_c$ and standardized household residuals

$$(3) \quad e^*_{ch} = \frac{e_{ch}}{\widehat{\sigma}_{\varepsilon,ch}} - \left[\frac{1}{H}\Sigma_{ch}\frac{e_{ch}}{\widehat{\sigma}_{\varepsilon,ch}}\right],$$

respectively, where H is the number of observations. The second term in e^*_{ch} adjusts for weighting at the first stage. One can avoid making any specific distributional form assumptions by drawing directly from the standardized residuals. Alternatively, percentiles of the empirical distribution of the standardized residuals can be compared to the corresponding percentiles of standardized normal, t, or other distributions.

Before proceeding to simulation, the estimated variance-covariance matrix, $\widehat{\Sigma}$, weighted by the household expansion factors, ℓ_{ch}, is used to obtain GLS estimates of the first-stage

construct a Wald test for homoscedasticity (Greene (2000), Section 12.5.3). Allowing the bounds to be freely estimated generates a standardized distribution for predicted disturbances which is well behaved in our experience. This is particularly important when using the standardized residuals directly in a semi-parametric approach to simulation (see Section 7 below.) However, we have also found that imposing a minimum bound of zero and a maximum bound $A^* = (1.05)\max\{e^2_{ch}\}$ yields similar estimates of the parameters α .

parameters, $\widehat{\beta}_{GLS}$, and their variance, $Var(\widehat{\beta}_{GLS})$.[8] In our experience, model estimates have been very robust to estimation strategy, with weighted GLS estimates not significantly different from the results of OLS or quantile regressions weighted by expansion factors. The GLS estimates do not differ significantly from coefficients obtained from weighted quantile regressions.

4. The Welfare Estimator

Although disaggregation may be along any dimension - not necessarily geographic - for convenience we refer to our target populations as 'villages'. There are M_v households in village v and household h has m_h family members. To study the properties of our welfare estimator as a function of population size we assume that the characteristics x_h and the family size m_h of each household are drawn independently from a village-specific constant distribution function $G_v(x, m)$: the super population approach.

While the unit of observation for expenditure in these data is typically the household, we are more often interested in poverty and inequality measures based on individuals. Thus we write $W(m_v, X_v, \beta, u_v)$, where m_v is an M_v-vector of household sizes in village v, X_v is a $M_v \times k$ matrix of observable characteristics and u_v is an M_v-vector of disturbances.

[8] Consider the GLS model

$$y^* = X^*\beta + \varepsilon^*,$$

where $y^* = Py$, etc. $\mathrm{E}[\varepsilon\varepsilon^T] = \Omega$, W is a weighting matrix of expansion factors, and $P^TP = W\Omega^{-1}$. Then $Var(\widehat{\beta}_{GLS}) = (X^TW\Omega^{-1}X)^{-1}(X^TW\Omega^{-1}WX)(X^TW\Omega^{-1}X)^{-1}$.

Because the vector of disturbances for the target population, u_v, is unknown, we estimate the expected value of the indicator given the village households' observable characteristics and the model of expenditure. This expectation is denoted $\mu_v = E[W|m_v, X_v, \zeta_v]$, where ζ_v is the vector of model parameters, including those which describe the distribution of the disturbances. For most poverty measures W can be written as an additively separable function of household poverty rates, $w(x_h, \beta, u_h)$, and μ_v can be written

$$(4) \quad \mu_v = \frac{1}{N_v} \sum_{h \in H_v} m_h \int_{u_h} w_h(x_h, \beta, u_h) d\mathcal{F}^{vh}(u_h),$$

where H_v is the set of all households in village v, $N_v = \sum_{h \in H_v} m_h$ is the total number of individuals, and $\mathcal{F}^{vh}$ is the marginal distribution of the disturbance term of household h in village v. When W is an inequality measure, however, the contribution of one household depends on the level of well-being of other households and W is no longer separable. Then we need the more general form,

$$(5) \quad \mu_v = \int_{u_1} \cdots \int_{u_{M_v}} W(m_v, X_v, \beta, u_v) \, d\mathcal{F}^v(u_{M_v}, \ldots, u_1),$$

where $u_1 \ldots u_{M_v}$ are the disturbance terms for the M_v households in village v.

In constructing an estimator of μ_v we replace ζ_v with consistent estimators, $\widehat{\zeta}_v$, from the first stage expenditure regression. This yields $\widehat{\mu}_v = \mathrm{E}[W \mid m_v, X_v, \widehat{\zeta}_v]$. This expectation is often analytically intractable so simulation or numerical integration are used to obtain the estimator $\widetilde{\mu}_v$.

5. Properties and Precision of the Estimator

The difference between $\widetilde{\mu}$, our estimator of the expected value of W for the village, and the actual level may be written

$$(6) \quad W - \widetilde{\mu} = (W - \mu) + (\mu - \widehat{\mu}) + (\widehat{\mu} - \widetilde{\mu}).$$

(The index v is suppressed here and below). Thus the prediction error has three components: the first due to the presence of a disturbance term in the first-stage model which causes households' actual expenditures to deviate from their expected values (idiosyncratic error); the second due to variance in the first-stage estimates of the parameters of the expenditure model (model error); and the last due to using an inexact method to compute $\widehat{\mu}$ (computation error). The error components are uncorrelated (see below). We consider the properties of each:[9]

Idiosyncratic Error - $(W - \mu)$

The actual value of the welfare indicator for a village deviates from its expected value, μ, as a result of the realizations of the unobserved component of expenditure in that village. Figure 2 illustrates. For convenience, denote the known expenditure component $\{\mathbf{x}_h^T\beta\}$ as t_h. Randomly drawn vectors $\mathbf{u}^r$ are added to $\mathbf{t}$ and empirical distributions of log

[9]Our target is the level of welfare that could be calculated if we were fortunate enough to have *observations* on expenditure for all households in a population. Clearly because expenditures are measured with error this may differ from a measure based on true expenditures. See Chesher and Schluter (2002) for methods to estimate the sensitivity of welfare measures to mismeasurement in y.

per-capita expenditure are graphed. The first panel shows the cumulative distribution of log per-capita expenditure based on a single simulation draw for 10 households. Subsequent panels superimpose 25 simulations for target populations of increasing size (where, for the purpose of illustration, u_h is assumed to be distributed *iid* $\mathcal{N}(0, \sigma^2)$). For small populations there is considerable variation in distributions across realizations of $\mathbf{u}$. It is easily proved that a limiting picture, that is for an infinite-sized population, will portray the underlying distribution. As is clear from Figure 2, particular realizations of $\mathbf{u}$ lose their effect on the empirical distribution of consumption.

When W is separable, this error is a weighted sum of household contributions:

$$(7) \quad (W - \mu) = \frac{1}{\overline{m}_M} \frac{1}{M} \sum_{h \in H_v} m_h \left[w(x_h, \beta, u_h) - \int_{u_h} w(x_h, \beta, u_h) d\mathcal{F}(u_h) \right],$$

where $\overline{m}_M = N/M$ is the mean household size among M village households. As the village population size increases, new values of x, and m are drawn from the constant distribution function $G_v(x, m)$. To draw new error terms in accordance with the model $u_{ch} = \eta_c + \varepsilon_{ch}$ complete enumeration areas are added, independently of previous EAs. Since $\overline{m}_M$ converges in probability to $\mathrm{E}[m]$,

$$(8) \quad \sqrt{M}(\mu - W) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\mathrm{I}}) \quad \text{as } M \to \infty,$$

where

$$(9) \quad \Sigma_{\mathrm{I}} = \frac{1}{(\mathrm{E}[m])^2} \mathrm{E}[m_h^2 \mathrm{Var}(w | x_h, \beta)].$$

When W is a non-separable inequality measure there usually is some pair of functions f and g, such that W may be written $W = f(\overline{y}, \overline{g})$,where $\overline{y} = \frac{1}{N}\sum_{h\in H_v} m_h y_h$ and $\overline{g} = \frac{1}{N}\sum_{h\in H_v} m_h g(y_h)$ are means of independent random variables.[10] The latter may be written

$$(10)\quad \overline{g} = \frac{1}{\overline{m}_M}\frac{1}{M}\sum_{h\in H_v} m_h g(y_h),$$

which is the ratio of means of M *iid* random variables $g_h = m_h g(y_h)$ and m_h. Assuming that the second moments of g_h exist, $\overline{g}$ converges to its expectation and is asymptotically normal. The same remark holds for $\overline{y}$. Thus, non-separable measures of welfare also converge as in (8) for some covariance matrix Σ_{I}.[11]

The idiosyncratic component, $\mathrm{V_I} = \Sigma_{\mathrm{I}}/M$, falls approximately proportionately in M. Said conversely, this component of the error in our estimator increases as one focuses on smaller target populations, which limits the degree of disaggregation possible. At what population size this error becomes unacceptably large depends on the explanatory power of the x variables in the expenditure model and, correspondingly, the importance of the remaining idiosyncratic component of expenditure.

Model Error - $(\mu - \widehat{\mu})$

[10]The Gini coefficient is an exception but it can be handled effectively with a separable approximation. See Elbers, *et. al.* (2000)

[11]The above discussion concerns the asymptotic properties of the welfare estimator, in particular consistency. In practice we simulate the idiosyncratic variance for an actual sub-population rather than calculate the asymptotic variance.

This is the second term in the error decomposition of equation (6). The expected welfare estimator $\widehat{\mu} = \mathrm{E}[W \mid m_v, X_v, \widehat{\zeta}_v]$ is a continuous and differentiable function of $\widehat{\zeta}$, which are consistent estimators of the parameters. Thus $\widehat{\mu}$ is a consistent estimator of μ and:

$$(11) \quad \sqrt{s}(\mu - \widehat{\mu}) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\mathrm{M}}) \quad \text{as } s \to \infty,$$

where s is the number of survey households used in estimation.[12] We use the delta method to calculate the variance Σ_{M}, taking advantage of the fact that μ admits of continuous first-order partial derivatives with respect to ζ. Let $\bigtriangledown = [\partial\mu \ / \partial\zeta]|_{\widehat{\zeta}}$ be a consistent estimator of the derivative vector. Then $\mathrm{V}_{\mathrm{M}} = \Sigma_{\mathrm{M}}/s \approx \bigtriangledown^T \mathrm{V}(\widehat{\zeta}) \bigtriangledown$, where $\mathrm{V}(\widehat{\zeta})$ is the asymptotic variance-covariance matrix of the first stage parameter estimators.

Because this component of the prediction error is determined by the properties of the first stage estimators, it does not increase or fall systematically as the size of the target population changes. Its magnitude depends, in general, only on the precision of the first-stage coefficients and the sensitivity of the indicator to deviations in household expenditure. For a given village v its magnitude will also depend on the distance of the explanatory x variables for households in that village from the levels of those variables in the sample data.

[12] Although $\widehat{\mu}$ is a consistent estimator, it is biased. Our own experiments and analysis by Saul Morris (IFPRI) for Honduras indicate that the degree of bias is extremely small. We thank him for his communication on this point. Below we suggest using simulation to integrate over the model parameter estimates, $\widehat{\zeta}$, which yields an unbiased estimator.

Computation Error - $(\widehat{\mu} - \widetilde{\mu})$

The distribution of this component of the prediction error depends on the method of computation used. When simulation is used this error has the asymptotic distribution given below in (16). It can be made as small as computational resources allow.

The computation error is uncorrelated with the model and idiosyncratic errors. There may be some correlation between the model error, caused by disturbances in the sample survey data, and the idiosyncratic error, caused by disturbances in the census, because of overlap in the samples. However, the approach described here is necessary precisely *because* the number of sampled households that are also part of the target population is very small. Thus, we can safely neglect such correlation.

For two populations, say Q and K, one can test whether the difference in their expected welfare estimates is statistically significant using the statistic

$$(12) \quad \frac{(\widetilde{\mu}_Q - \widetilde{\mu}_K)^2}{\mathrm{Var}[(\widetilde{\mu}_Q - W_Q) - (\widetilde{\mu}_K - W_K)]},$$

which is distributed asymptotically $\chi^2(1)$ under the null hypothesis $\mathrm{H}_0 : W_Q = W_K$. The parts of the variance in the prediction error for populations Q and K due to computation and the idiosyncratic component of W are independent. However, if the same first-stage model estimates are used to estimate t_h for households in both populations, then the model component of the prediction error will be correlated across populations. Let ψ be

a vector of all of the parameters used in the estimation of either $\widetilde{\mu}_Q$ or $\widetilde{\mu}_K$, and let $\mathbf{q}$ be a vector of the partial derivatives $[\partial(\widetilde{\mu}_Q - \widetilde{\mu}_K)/\partial\psi]|\widehat{\ }$. Then,

$$(13)\quad \mathrm{Var}[(\widetilde{\mu}_Q - W_K) - (\widetilde{\mu}_Q - W_K)] \approx \mathbf{q}^T \mathrm{V}\left(\widehat{\psi}\right)\mathbf{q} + \mathrm{V}_{\mathrm{I}}^Q + \mathrm{V}_{\mathrm{I}}^K + \mathrm{V}_{\mathrm{C}}^Q + \mathrm{V}_{\mathrm{C}}^K.$$

If the first-stage parameter estimates used to estimate household expenditure differ across the two regions then the first term is simply $\mathrm{V}_{\mathrm{M}}^Q + \mathrm{V}_{\mathrm{M}}^K$.

6. Computation

We use Monte Carlo simulation to calculate: $\widehat{\mu}$, the expected value of the welfare measure given the first stage model of expenditure; V_{I}, the variance in W due to the idiosyncratic component of household expenditures; and the gradient vector $\nabla = [\partial\mu/\partial\zeta]|_{\widehat{\zeta}}$.

Let the vector $\widehat{u}^r$ be the r^{th} simulated disturbance vector. Treated parametrically, $\widehat{u}^r$ is constructed by taking a random draw from an M_v-variate standardized distribution and pre-multiplying this vector by a matrix T, defined such that $TT^T = \widehat{\Sigma}$. Treated semi-parametrically, $\widehat{u}^r$ is drawn from the residuals with an adjustment for heteroscedasticity. We consider two approaches. First, a location effect, $\widehat{\eta}_c^r$, is drawn randomly, and with replacement, from the set of all sample $\widehat{\eta}_c$. Then an idiosyncratic component, e_{ch}^{*r}, is drawn for each household κ with replacement from the set of all standardized residuals and $e_{c\kappa}^r = \widehat{\sigma}_{\varepsilon,c\kappa}(e_{ch}^{*r})$. The second approach differs in that this component is drawn only from the standardized residuals e_{ch}^* that correspond to the cluster from which household

κ's location effect was derived. Although $\widehat{\eta}_c$ and e_{ch} are uncorrelated, the second approach allows for non-linear relationships between location and household unobservables. It is considered empirically in the example below, Section 7.

With each vector of simulated disturbances we construct a value for the indicator, $\widehat{W}_r = W(m, \widehat{t}, \widehat{u}^r)$, where $\widehat{t} = \mathrm{X}^T\widehat{\beta}$, the predicted part of log per-capita expenditure. The simulated expected value for the indicator is the mean over R replications,

$$(14)\quad \widetilde{\mu} = \frac{1}{R}\sum_{r=1}^{R}\widehat{W}_r.$$

The variance of W around its expected value μ due to the idiosyncratic component of expenditures can be estimated in a straightforward manner using the same simulated values,

$$(15)\quad \widetilde{\mathrm{V}}_{\mathrm{I}} = \frac{1}{R}\sum_{r=1}^{R}(\widehat{W}_r - \widetilde{\mu})^2.$$

Simulated numerical gradient estimators are constructed as follows: We make a positive perturbation to a parameter estimate, say $\widehat{\beta}_k$, by adding $\delta|\widehat{\beta}_k|$, and then calculate $\widehat{t}^+$, followed by $\widehat{W}_r^+ = W(m, \widehat{t}^+, \widehat{u}^r)$, and $\widetilde{\mu}^+$. A negative perturbation of the same size is used to obtain $\widetilde{\mu}^-$. The simulated central distance estimator of the derivative $\partial\mu/\partial\beta_k|\sim$ is $(\widetilde{\mu}^+ - \widetilde{\mu}^-)/(2\delta|\widehat{\beta}_k|)$. As we use the same simulation draws in the calculation of $\widetilde{\mu}$, $\widetilde{\mu}^+$and $\widetilde{\mu}^-$, these gradient estimators are consistent as long as δ is specified to fall sufficiently rapidly as $R \to \infty$ (Pakes and Pollard (1989)). Having thus derived an estimate of the

gradient vector $\nabla = [\partial\mu/\partial\zeta]|_{\widehat{\zeta}}$, we can calculate $\widetilde{\mathrm{V}}_{\mathrm{M}} = \nabla^T \mathrm{V}(\widehat{\zeta})\nabla$.

Because $\widetilde{\mu}$ is a sample mean of R independent random draws from the distribution of $(W \mid m, \widehat{t}, \widehat{\Sigma})$, the central limit theorem implies that

$$(16)\quad \sqrt{R}(\widetilde{\mu} - \widehat{\mu}) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\mathrm{C}}) \quad \text{as } R \to \infty,$$

where $\Sigma_{\mathrm{C}} = \mathrm{Var}(W|m, \widehat{t}, \widehat{\Sigma})$.[13]

When the decomposition of the prediction error into its component parts is not important, a far more efficient computational strategy is available. Write

$$\ln y_{ch} = x_{ch}^T \beta + \eta_c(\zeta) + \varepsilon_{ch}(\zeta),$$

where we have stressed that the distribution of η and ε depend on the parameter vector ζ. By simulating ζ from the sampling distribution of $\widehat{\zeta}$, and $\{\eta_c^r\}$ and $\{\varepsilon_{ch}^r\}$ conditional on the simulated value ζ^r, we obtain simulated values $\{y_{ch}^r\}$, consistent with the model's distributional characteristics, from which welfare estimates W^r can be derived (Mackay (1998)). Estimates of expected welfare, μ, and its variance are calculated as in equations (14) and (15). Drawing from the sampling distribution of the parameters replaces the delta method as a way to incorporate model error into the total prediction error. Equation (15) now gives a sum of the variance components $\widetilde{\mathrm{V}}_{\mathrm{I}} + \widetilde{\mathrm{V}}_{\mathrm{M}}$, while Σ_{C} in equation (16)

[13]Whenever a parametric distribution is used, efficiency can be improved using a minimum discrepancy estimator, where draws are made systematically from the disturbance distribution (see Traub and Werschulz, 1998). In experiments estimating the headcount measure, we found that, for $R < 100$, $\sqrt{\mathrm{V}_{\mathrm{C}}}$ for this estimator was 74-78% of its value for Monte Carlo simulation.

becomes $\Sigma_C = \text{Var}(W|m, X, \widehat{\zeta}, \text{V}(\widehat{\zeta}))$.

7. Basic Simulation Results

This section uses the 1994 Ecuadorian *Encuesta Sobre Las Condiciones de Vida,* a household survey following the general format of a World Bank Living Standards Measurement Survey. It is stratified by 8 regions and intended to be representative at that level. Within each region there are several levels of clustering. At the final level, 12 to 24 households are randomly selected from a census enumeration area. Expansion factors allow the calculation of regional totals. The analysis in this section uses data from the rural Costa region.

Table 1 gives diagnostics for four different first-stage regressions. The first column refers to a regression with a range of demographic and education variables, but excluding all information about infrastructure. The second column corresponds to a regression where regressors include means of some of these same variables. The third column has results for a model with no means but including household level infrastructure variables, and the last column corresponds to a 'full' model with regressors chosen from all household level variables and also some of their means.[14] Detailed results for the full model are

[14]In order to choose which variable means to include we first estimated the model with only household level variables. We then estimated the residual location effect for each cluster in rural Costa, and regressed them on variable means to determine a set of means particularly suited to explaining the effect of location. We limited the chosen number of variables to five so as to avoid over-fitting our 39 sample cluster effects.

presented in Appendix 2, Table A.1.

All of the regressions are weighted by population expansion factors. These weights differ considerably across clusters and the test results in row one of Table 1 indicate that weighting has a significant effect on the coefficients. Weighting is discussed further in subsection below.

In row 3 we examine the varying importance of residual intra-cluster correlation across the different models by decomposing the overall disturbance variance. The (weighted) cluster random effect variance, σ_η^2, is estimated non-parametrically, allowing for heteroscedasticity in ε_{ch}. For details, along with the formula used to estimate $\text{Var}(\widehat{\sigma}_\eta^2)$, see Appendix 1. Further evidence on the importance of residual location effects is provided by a regression of the total residuals, $\widehat{u}$, on cluster fixed effects. Row 4 gives results of an F-test of the null hypothesis that fixed effect coefficients are jointly zero. Both rows 3 and 4 indicate that there is a significant intra-cluster correlation in the disturbances of models that do not include location mean variables. However, when means of household-level variables are included as regressors they effectively capture most of the effect of location on consumption. Infrastructure variables also contribute, and in the full model there is little remaining evidence of spatial correlation in the residuals.

We next model the variance of the idiosyncratic part of the disturbance, $\sigma_{\varepsilon,ch}^2$. In Section 3 we suggested estimating a logistic model with free bounds. However, we have found

that imposing a minimum bound of zero and a maximum bound $A^* = (1.05)\max\{e_{ch}^2\}$ yields similar estimates of the parameters α . These restrictions allow one to estimate the simpler form:

$$ln\left[\frac{e_{ch}^2}{A^* - e_{ch}^2}\right] = z_{ch}^T\alpha + r_{ch},$$

which is what we do here.[15] Detailed results corresponding to the full model may again be found in Appendix 2, Table A.2. Results of chi-square tests of the null that estimated parameters are jointly zero in these regressions are found in row 5 of Table 1, where homoscedasticity is clearly rejected for all but the first model specification. Letting $\exp\{\mathbf{z}_{ch}^T\widehat{\boldsymbol{\alpha}}\} = B$ and using the delta method, the model implies a household specific variance estimator for ε_{ch} of

$$(17)\quad \widehat{\sigma}_{\varepsilon,ch}^2 = \left[\frac{AB}{1+B}\right] + \frac{1}{2}\widehat{\mathrm{Var}}(r)\left[\frac{AB(1-B)}{(1+B)^3}\right].$$

Finally, the last rows in Table 1 present results of tests of the null hypotheses that η and ε are distributed normally, based on the cluster residuals $\widehat{\eta}_c$ and standardized household residuals e_{ch}^*, respectively.

For some strata in Ecuador the standarized residual distribution appears to be approximately normal, even if formally rejected by tests based on skewness and kurtosis.

[15]Specifying the bounds is problematic in that it generates some small values of $\widehat{\sigma}_{\varepsilon,ch}$ and, consequently, very large absolute standardized residuals. Thus, when simulating on the basis of the empirical distribution of these residuals we drop four observations with $e^* > |5|$.

Elsewhere, we find a $t(5)$ distribution to be the better approximation. Relaxing the distributional form restrictions on the disturbance term and taking either of the semi-parametric approaches outlined above makes very little difference in the results for our Ecuadorian example.

Simulation results for the headcount measure of poverty and the general entropy (0.5) measure of inequality are in Tables 2 and 3. We construct populations of increasing size from a constant distribution $G_v(x, m)$ by drawing households randomly from all census households in the rural Costa region. They are allocated in groups of 100 to pseudo enumeration areas, with '*parroquias*' of a thousand households created out of groups of ten EAs. We continue aggregating to obtain nested populations with 100 to 100,000 households.

For each model and measure we present estimates of the expected value of the welfare indicator, calculated with a sufficient number of simulation draws to ensure that the standard error due to computation is less than 0.001. In all examples we adjust for outliers. In standard situations, where the analyst has direct information about y, it is common to have outliers in that variable due to mismeasurement, inputting errors, etc. The problem is typically dealt with by discarding suspect observations. Here we have an analoguous problem with respect to the $\mathbf{x}$ variables used to infer expenditure levels, and

we deal with it in the usual way.[16] In addition to the standard "dirty data" problem, when treating the distribution of u_h parametrically there is a non-zero probability of getting an extreme simulation draw and therefore an 'outlying' value for y_h. This problem is resolved by using truncated distributions. Since it is the best information we have, we use the minimum and maximum of $\widehat{\eta}_c$ and $\widehat{u}_{ch}$ from our first-stage log-expenditure regression as truncation points.[17] Poverty measures give zero weight to expenditure levels above the poverty line and are not very sensitive to variations below. Inequality measures, however, can be very sensitive to outlying values and therefore the choices made to discard observations and 'trim' disturbances. (Sampling raises similar issues and this subject is an area of continuing research.)

Table 2, column 1, refers to the headcount measure of poverty. It is defined as

$$(18)\quad W = \frac{1}{N}\sum_{h\in H_v} m_h \mathbb{1}(y_h < z),$$

where z is a poverty line defined in per-capita expenditure terms and $\mathbb{1}(\,)$ is an indicator function taking on the value of one if the expression inside of the brackets is true and zero otherwise. When η_c and ε_{ch} are normally distributed there is a simple analytical form for

[16] We delete households with predicted per-capita expenditure, $\widehat{t}_h$, outside the range of observed per-capita expenditure in the household survey, losing less than 0.2% of our total census observations as a result.

[17] Although they are in line with common practice, both steps of this procedure are admittedly somewhat *ad hoc*. Addressing the standard problem of mismeasurement in y_h, Cowell and Victoria-Feser (1996) suggest leaving suspected outliers in the data when estimating inequality and using weighting to lessen their importance. A similar approach could be taken here.

the welfare estimator:

$$(19)\quad \widehat{\mu} = \frac{1}{N}\sum_{h\in H_v} m_h \Phi((\ln z - \widehat{t}_h)/\widehat{\sigma}_h),$$

where $\Phi(.)$ is the standard normal distribution function and $\widehat{\sigma}_h = \sqrt{\widehat{\sigma}^2_\eta + \widehat{\sigma}^2_{\epsilon,ch}}$. Table 2, column 2, refers to the general entropy (GE) measure with parameter $c = 0.5$. This measure is defined as

$$(20)\quad W_c = \frac{1}{c(1-c)}\left\{1 - \frac{1}{N}\sum_{h\in H_v} m_h (\frac{y_h}{\overline{y}})^c\right\}.$$

The first set of results (I) is calculated using the full first-stage model (column four of Table 1). Here we assume that the location effect estimated at the cluster level in the survey data applies in the census to an enumeration area, and that household disturbances across different EAs are uncorrelated. The set of results (II) again are calculated using the full first-stage model, but now with the (conservative) assumption that the location effect estimated from clusters applies across an entire *parroquia*. This has the expected effect of increasing the idiosyncratic variance, although the estimator is still remarkably good given the small size of the residual location effect once infrastructure means are included as observable correlates of consumption. For comparison, (III) and (IV) give simulation results using the most sparse first-stage model - that with only household-level variables and no means (column one of Table 1). In (III) we estimate μ as in (I), with the location effect at the EA level, while in (IV) we impose the assumption that there

is no intra-cluster correlation, i.e. that $\eta = 0$. A comparison of the results in (I) and (III) highlights the importance of developing a set of regressors that succeeds in picking up most of the influence of location on consumption. The prediction errors in (III) are higher, particularly for inequality. As noted above, there is great potential to enrich both the survey and census with other data to obtain appropriate variables. Comparing (III) and (IV) one sees that failing to allow for the effect of location can lead to a markedly over-optimistic view of the precision of the estimator.

Table 3 shows estimates of the expected value of the welfare indicator, the standard error of the prediction, and the share of the total variance due to the idiosyncratic component for increasingly large target populations. The location effect estimated at the cluster level in the survey data is applied to EAs in the census. In all cases the standard error due to computation is less than 0.001.

Looking across columns one sees how the variance of the estimator falls as the size of the target population increases. For both measures the total standard error of the prediction falls to about five to seven percent of the point estimate with a population of just 15,000 households. At this point, the share of the total variance due to the idiosyncratic component of expenditure is already small, so there is little to gain from moving to higher levels of aggregation. The table also shows that estimates for populations of 100 have large errors Clearly it would be ill advised to use this approach to determine the poverty

of yet smaller groups or single households.

We now examine briefly several other modeling choices. First we consider the importance of modelling heteroscedasticity in the idiosyncratic component of the disturbance. We estimate expected headcount and GE (0.5) measures for the entire rural Costa, by *parroquia*, first using a model of heteroscedasticity and then assuming homoscedasticity. Table 4, column 1, indicates that there is little re-ranking of *parroquia*s based on their headcount measures when heteroscedasticity is ignored. However, allowance for heteroscedasticity does have an important effect on rankings by inequality. The bottom half of the table indicates that the Spearman's rank correlation of general entropy inequality estimates is just 0.83. The difference in estimates within each *parroquia* is not always trivial for either measure. Differences across the two sets of estimates reach 0.08 and 0.11 for the headcount and GE (0.5) measure, respectively.

We next consider the effect of weighting by population expansion factors. As noted above, all of our analyses use these weights. The argument for doing so is that there may be some variance in the parameters ζ within regions which is not modelled. If so, because we want to use the model estimates to impute into the census, we would prefer the model to fit most closely the clusters that represent large census populations. However, this decision is not innocuous. The expansion factors range by a factor of about 600, with about half of the clusters receiving on the order of 100 times as much weight in the

regression as the other half. To explore this, we estimate *parroquia* welfare measures using the full first-stage model without weighting by population expansion factors. Column two of Table 4 shows that this choice is very important. The rank correlation across weighted and unweighted estimates of the expected headcount is just 0.77, the average absolute difference is 0.05, and reaches as high as 0.34. For the general entropy measure, the rank correlation is similar: 0.78, with a maximum difference of 0.19.

Finally, we consider the second of the semi-parametric approaches to estimating the effect of the unobserved component of consumption on the welfare measure (see Section 6). Results are found in the third column of Table 4. Relaxing the functional form restrictions on the disturbance term makes very little difference in this example. The rank correlations between the parametric and semi-parametric treatments is 1.00 and 0.98 for the headcount and GE (0.5) measure, respectively, with maximum differences in the estimates of 0.04 and 0.05.

8. How much Improvement?

Most users of welfare indicators rely, by necessity, on sample survey based estimates. Table 5 demonstrates how much is gained by combining data sources. The second column gives the sampling errors on headcount measures estimated for each stratum using the survey data alone (taking account of sample design). There is only one estimate per

region as this is the lowest level at which the sample is representative. The population of each region is in the third column. When combining census and survey data it becomes possible to disaggregate to sub-regions and estimate poverty for specific localities. Here we choose as sub-regions *parroquias* or, in the cities of Quito and Guayaquil, *zonas*, because our prediction errors for these administrative units are similar in magnitude to the survey based sampling error on the region level estimates. (See the median standard error among sub-regions in the fourth column.) The final column gives the median population among these sub-regions. Comparing the third and final columns it is clear that, for the same prediction error commonly encountered in sample data, one can estimate poverty using combined data for sub-populations of a hundredth the size. This becomes increasingly useful the more there is spatial variation in well-being that can be identified using this approach. Considering this question, Demombynes, et. al. (2002) find, for several countries, that most sub-region headcount estimates do differ significantly from their region's average level.

9. Other Measures

Table 6 summarizes results for a range of welfare measures, again using the four nested census populations described above. In each case, location effects are assumed to apply

at the EA level. The measures are the FGT (1) measure of the severity of poverty,

$$(21)\quad W_1 = \frac{1}{N}\sum_{h\in H_v} m_h(1-\frac{y_h}{z})\mathbb{1}(y_h < z);$$

the variance of log expenditure,

$$(22)\quad W = \frac{1}{N}\sum_{h\in H_v} m_h(\ln y_h - \overline{\ln y})^2;$$

and the Atkinson measure with inequality aversion parameter of 2,

$$(23)\quad W_2 = 1 - \left\{\frac{1}{N}\sum_{h\in H_v} m_h(\frac{y_h}{\overline{y}})^{-1}\right\}^{-1},$$

where the village mean expenditure, $\overline{y}$, is weighted by household size.

Results for the FGT(1) measure, often called the poverty gap, are similar to those for the headcount. Again quite precise estimates are obtained for populations of just 15,000 households. Results for the variance of log expenditure measure are similar to those for the GE (0.5) measure presented in Table 3. Our estimates of the Atkinson measure are somewhat more precise that the other inequality measures.

10. Putting the Indicators to Work – Illustrations

We now use estimates of distributional measures in two different types of applications. The measures have been calculated for all *parroquias* in rural Ecuador using the full census. *Parroquias* are the lowest adminstrative units. The calculations are based on

three separate regional first-stage consumption models (estimation results available from the authors on request).

Geographical Maps of Welfare

A useful way of understanding the geographical spread of poverty or inequality is to contruct a map using GIS data. Figure 3 provides an example. Comparisons between the Costa, the coastal region of Ecuador, and the Sierra, the central mountainous region, feature highly in popular political debate in Ecuador.[18] The top two maps in Figure 3 depict the spatial distribution of poverty on the basis of two common measures: the headcount and the poverty gap, FGT(1).[19] The bottom two maps in Figure 3 indicate those instances where the two alternative poverty measures differ in their ranking of cantons. The map on the lower left shows that in the Costa a number of cantons are ranked poorer under the headcount criterion than under the poverty gap. In contrast, in the Sierra, numerous cantons are ranked more poor under the poverty gap criterion than under the headcount. Clearly, views about the relative poverty of the regions will be affected by the measure of poverty employed. It is also clear that, irrespective of the poverty measure used, all cantons in the eastern part of Ecuador are particularly poor.

This type of map could be used for targetting development efforts, or for exploring relationships between welfare indicators and other variables. For example, a poverty or

[18]See, for example, "Under the Volcano", *The Economist*, November 27, 1999, p. 66.

[19]For visibility we have disaggregated only to *cantons*, the administrative level just above a *parroquia*.

inequality map could be overlaid with maps of other types of data, say on agro-climatic or other environmental characteristics. The visual nature of the maps may highlight unexpected relationships that would escape notice in a standard regression analysis.

Are Neighbors Equal?

An important issue in the area of political economy and public policy is to determine the appropriate level of government to give responsibility for public services and their financing. The advantage of decentralizing to make use of better community-level information about priorities and the characteristics of residents may be offset by a greater likelihood that the local governing body is controlled by elites - to the detriment of weaker community members. In a recent paper, Bardhan and Mookherjee (1999) highlight the roles of both the level and heterogeneity of local inequality (and poverty) as determinants of the relative likelihood of capture at different levels of government. As most of the theoretical predictions are ambiguous, they stress the need for empirical research into the causes of political capture - analysis which has been held back by a lack of empirical measures for most variables.[20] Our community-level welfare estimates can help to address this problem.

We can answer, first, many questions about the level and heterogeneity of welfare

[20]Galasso and Ravallion (2002), which compares the inter- vs intra-district targetting of schooling in Bangladesh, uses village-level inequality measures, but is limited to those sampled in the household expenditure survey.

at different levels of government. For example, here we decompose inequality in rural Ecuador into between- and within-group components and examine how within-group inequality evolves at progressively lower levels of regional disaggregation. At one extreme, when a country-level perspective is taken, all inequality is, by definition, within-group. At the other extreme, when each individual household is taken as a separate group, the within-group contribution to overall inequality is zero (assuming, as is implicit in our use of a per-capita indicator, an equal distribution within each household). But how rapidly does the within-group share fall? Is it reasonable to suppose that at a sufficiently low level of disaggregation (say, a village or neighbourhood) differences within groups are small, and most of overall inequality is due to differences between groups?

We employ the general entropy (0.5) inequality measure because it is decomposable. If N individuals are placed in one of J groups subscripted by j, and the proportion of the population in the jth group, denoted f_j, has weighted mean per-capita expenditure $\overline{y}_j$ and inequality ω_j, then

$$(24)\quad W_{0.5} = 4\left\{1 - \sum_{j=1}^{J} f_j(\frac{\overline{y}_j}{\overline{y}})^{0.5}\right\} + \sum_{j=1}^{J} \omega_j f_j(\frac{\overline{y}_j}{\overline{y}})^{0.5},$$

where the first term is the inequality between groups and the second is within groups (Cowell, 1995). In stages we disaggregate the country down to the *parroquia* level. Table 7 illustrates that even at a very high degree of spatial disaggregation, 86% of overall rural

inequality can still be attributed to differences within groups.[21] For further interpretation and examples from other countries, see Elbers, et. al. (2002).

Thus, as often suggested by anecdotal evidence, even within local communities there exists a considerable heterogeneity of living standards. In addition to affecting the likelihood of political capture, this may have implications for the feasibility of raising revenues locally, as well as for the extent to which residents of such communities can be viewed as having similar demands and priorities.

Put together with either survey data on attitudes towards government or on the allocation of public spending, disaggregated inequality estimates could be used to directly assess the influence of welfare distributions on the political process. We plan to explore this further in the context of the targetting of social fund programs.

11. Conclusions

In constructing disaggregated estimates of welfare we have explored a straightforward idea. We use detailed household survey data to estimate a model of per-capita expenditure and then use the resulting parameter estimates to weight the census-based characteristics of a target population in determining its expected welfare level. While others have taken weighted combinations of variables in the census to estimate household poverty, this merging of data sources has the advantage of yielding estimators with

[21] We have confined our attention to rural areas where there is no evidence of spatial autocorrelation in ε. Results using all of Ecuador were very similar.

clear interpretations via their link to household expenditure; which are mutually comparable; and, perhaps most importantly, which can be assessed for reliability using standard statistical theory.

What is quite remarkable is how well this method of estimating welfare measures can work in practice. In our examples using Ecuadorian data we find that estimates are often quite reliable for populations as small as 15,000 households, a 'town'. This is a very considerable improvement over the direct survey-based estimates, which are only consistent for areas encompassing hundreds of thousands of households.

Given these promising initial results there is also no reason to be passive consumers of existing data sets. Governments and surveying bodies can be encouraged to design both census and survey instruments to correspond more closely for this purpose.

So now that we have estimates of poverty and inequality in thousands of 'towns' or other groups, what can we do with them? The possibilities seem many and varied. For many questions, intra-regional cross-town analysis could considerably enrich the existing results of cross-country studies (see, Elbers and Lanjouw, 2001). At the micro-level increasing attention is being paid to ways in which welfare distributions within groups relate to socioeconomic and political outcomes. Of the resulting multitude of theories, most remain to be tested. Again, our findings regarding the level and heterogeneity of well-being at different levels of government, features which have been linked *in theory* to

political capture and the targetting of public resources, are just one illustration of what is possible. Merging these measures with data on crime, education, health, voting patterns, unemployment, and so on, will open up many promising avenues for further research.

Department of Economics, Vrije Universiteit, De Boelelaan 1105, 1081 HV Amsterdam, N.L.; celbers@feweb.vu.nl,

and

Department of Agriculture and Resource Economics, University of California at Berkeley, and the Brookings Institution, 1775 Massachusetts Avenue NW, Washington, DC, 20036, U.S.A.; jlanjouw@brook.edu,

and

The World Bank, 1818 H. Street, Washington, DC, 20433, U.S.A.; planjouw@worldbank.org.

REFERENCES

AGHION, P., AND P. BOLTON (1997): "A Theory of Trickle Down Growth and Development," *Review of Economic Studies*, 64, 2, 151-72.

ALDERMAN, H., M. BABITA, G. DEMOMBYNES, N. MAKHATHA, AND B. ÖZLER (2002): "How Low Can You Go?: Combining Census and Survey Data for Mapping Poverty in South Africa," *Journal of African Economics,* forthcoming.

ALESINA, A., AND D. RODRIK (1994): "Distributive Politics and Economic Growth," *Quarterly Journal of Economics*, 109, 465-90.

ANGRIST, J. D., AND A.B. KRUEGER (1992): "The Effect of Age of School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association,* 87, 328-36.

ARELLANO, M., AND C. MEGHIR (1992): "Female Labour Supply and on the Job Search: an Empirical Model Estimated using Complementary Data Sets," *Review of Economic Studies,* 59, 537-59.

ATKINSON, A. B., AND A. BRANDOLINI (2001): "Promise and Pitfalls in the Use of "Secondary" Data-Sets: Income Inequality in OECD Countries," *Journal of Economic Literature*, 39, 3.

BANERJEE, A., AND E. DUFLO (2000): "Inequality and Growth: What Can the Data Say?," NBER Working paper no. 7793.

BANERJEE, A., AND A. NEWMAN (1993) "Occupational Choice and the Process of Development," *Journal of Political Economy*, 101, 1, 274-98.

BARDHAN, P., AND D. MOOKHERJEE (1999): "Relative Capture of Local and Central Governments: An Essay in the Political Economy of Decentralization," CIDER Working Paper no. C99-109, University of California at Berkeley.

BARRO, R., AND X. SALA-I-MARTIN (1991): "Convergence Across States and Regions," *Brookings Papers on Economic Activity*, no. 1, 107-82.

BRUNO, M., M. RAVAILLION, AND L. SQUIRE (1998): "Equity and Growth in Developing Countries: Old and New Perspectives on the Policy Issues," in *Income Distribution and High-Quality Growth,* eds. V. Tanzi and K.-Y. Chu. Cambridge: MIT Press.

CHESHER, A., AND C. SCHLUTER (2002): "Welfare Measurement and Measurement Error," *Review of Economic Studies,* forthcoming.

COWELL, F. (1995): *The Measurement of Inequality,* 2nd ed. Hemel Hempstead: Prentice Hall/Harvester Wheatsheaf.

COWELL, F., AND M.-P. VICTORIA-FESER (1996) "Robustness Properties of Inequality Measures," *Econometria,* 64, 1, 77-101.

DEATON, A. (1997): *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy.* Washington, D.C.: The Johns Hopkins University Press for the World Bank.

_____(1999): "Inequalities in Income and in Inequalities in Health," NBER Working paper no. 7141.

DEININGER, K., AND L. SQUIRE (1996): "A New Data Set Measuring Income Inequality," *The World Bank Economic Review,* 10, 565-91.

DEMOMBYNES, G., C. ELBERS, J. O. LANJOUW, P. LANJOUW, J. MISTIAEN, AND, B. Ö(2002): "Producing an Improved Geographic Profile of Poverty: Methodology and Evidence from Three Developing Countries," WIDER Discussion Paper no. 2002/39, The United Nations.

ELBERS, C., AND P. LANJOUW (2001): "Intersectoral Transfer, Growth, and Inequality in Rural Ecuador," *World Development,* 29, 3, 481-96.

ELBERS, C., J. O. LANJOUW, AND P. LANJOUW (2000): "Welfare in Villages and Towns: Micro-Measurement of Poverty and Inequality," Tinbergen Institute Working Paper no. 2000-029/2.

ELBERS, C., J. O. LANJOUW, P. LANJOUW, AND P. G. LEITE (2002): "Poverty and Inequality in Brazil: New Estimates from Combined PPV-PNAD Data," Unpublished Manuscript, The World Bank.

ELBERS, C., P. LANJOUW, J. MISTIAEN, B. ÖZLER, AND K. SIMLER (2002) : "Are Neighbours Equal? Estimating Local Inequality in Three Developing Countries," Unpublished Manuscript, The World Bank.

FIELDS, G. (1989): "A Compendium of Data on Inequality and Poverty for the Developing World," Unpublished Manuscript, Cornell University.

_____(2001): "Economic Growth and Inequality: A Review of the Empirical Evidence," Chapter 3 in *Distribution and Development: A New Look at the Developing World.* Russel Sage Foundation and MIT Press.

GALOR, O., AND J. ZEIRA (1993): "Income Distribution and Macroeconomics," *Review of Economic Studies*, 60, 35-52.

GALASSO, E., AND M. RAVALLION (2002): "Decentralized Targetting of an Anti-Poverty Program," Unpublished Manuscript, The World Bank.

GHOSH, M., AND J. N. K. RAO (1994): "Small Area Estimation: An Appraisal," *Statistical Science*, 9, 55-93.

GREENE, W. H. (2000): *Econometric Analysis.* Fourth Edition. New Jersey: Prentice-Hall Inc.

HELLERSTEIN, J., AND G. IMBENS (1999): "Imposing Moment Restrictions from Auxiliary Data by Weighting," *Review of Economics and Statistics,* 81, 1, 1-14.

KEYZER, M. (2000): "Reweighting Survey Observations by Monte Carlo Integration on a Census," Stichting Onderzoek Wereldvoedselvoorziening, Staff Working Paper no. 00.04, the Vrije Universiteit, Amsterdam.

LUSARDI, A. (1996): "Permanent Income, Current Income and Consumption: Evidence from Two Panel Data Sets," *Journal of Business and Economic Statistics*, 14, 1.

MACKAY, D. J. C. (1998): "Introduction to Monte Carlo Methods," in *Learning in Graphical Models; Proceedings of the NATO Advanced Study Institute,* ed. by M. I. Jordan. Kluwer Academic Publishers Group.

MURPHY, K. M., SHLEIFER, A., AND R.C. VISHNY (1989): "Income Distribution, Market Size and Industrialization," *Quarterly Journal of Economics*, 104, 537-64.

PERSSON, T., AND G. TABELLINI (1994): "Is Inequality Harmful for Growth," *American Economic Review*, 84, 600-21.

PAKES, A., AND D. POLLARD (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica,* 57, 1027-58.

RAO, J. N. K. (1999): "Some Recent Advances in Model-Based Small Area Estimation," *Survey Methodology*, 25, 175-86.

RAVALLION, M. (1998): "Does Aggregation Hide the Harmful Effects of Inequality on Growth?," *Economics Letters*, 61, 1, 73-7.

TAROZZI, A. (2002): "Estimating Comparable Poverty Counts from Incomparable Surveys: Measuring Poverty in India," RPDS Working paper no. 213, Princeton University.

TRAUB, J.F., AND A.G. WERSCHULZ (1998): *Complexity and Information.* Cambridge: Cambridge University Press.

FIGURES AND TABLES

Figure 1a

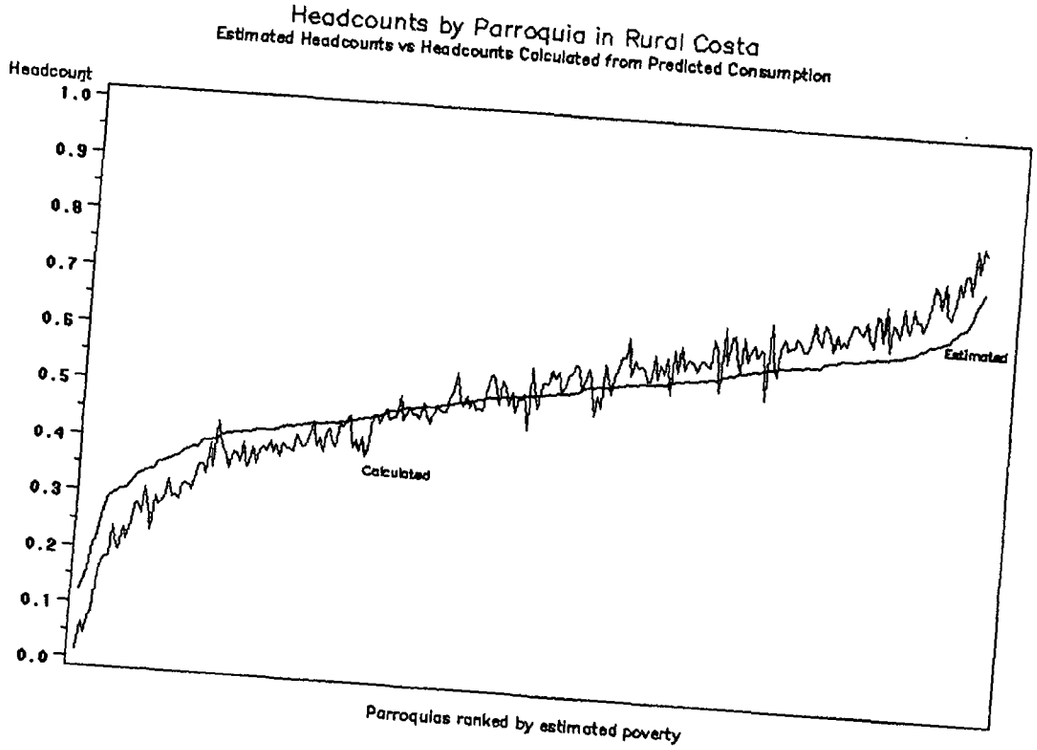


Figure 1b

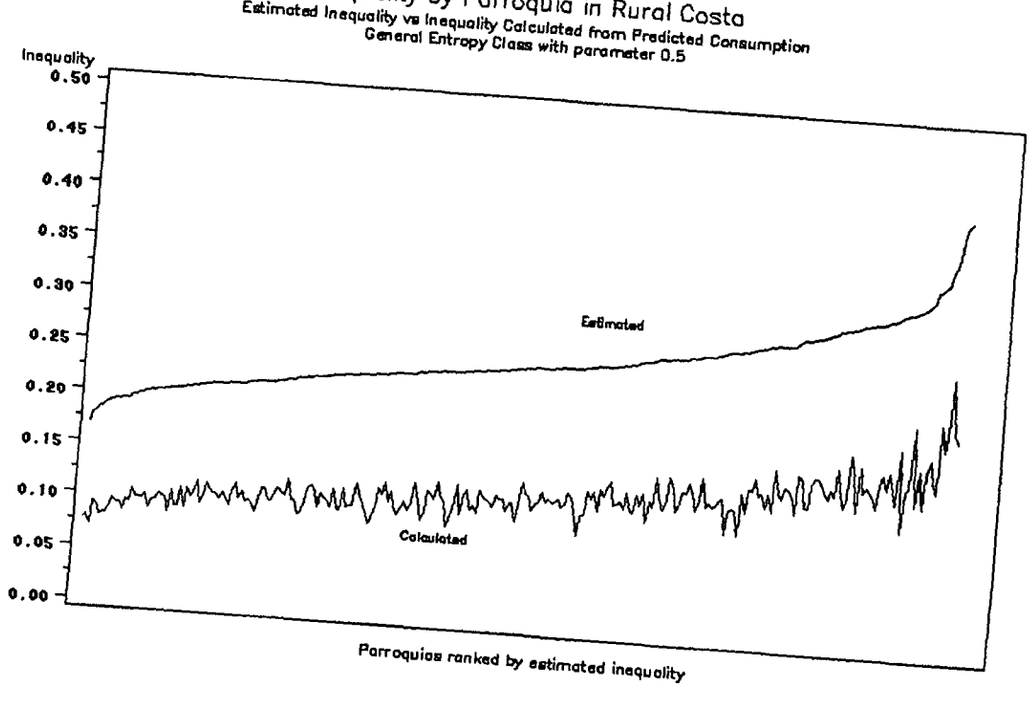
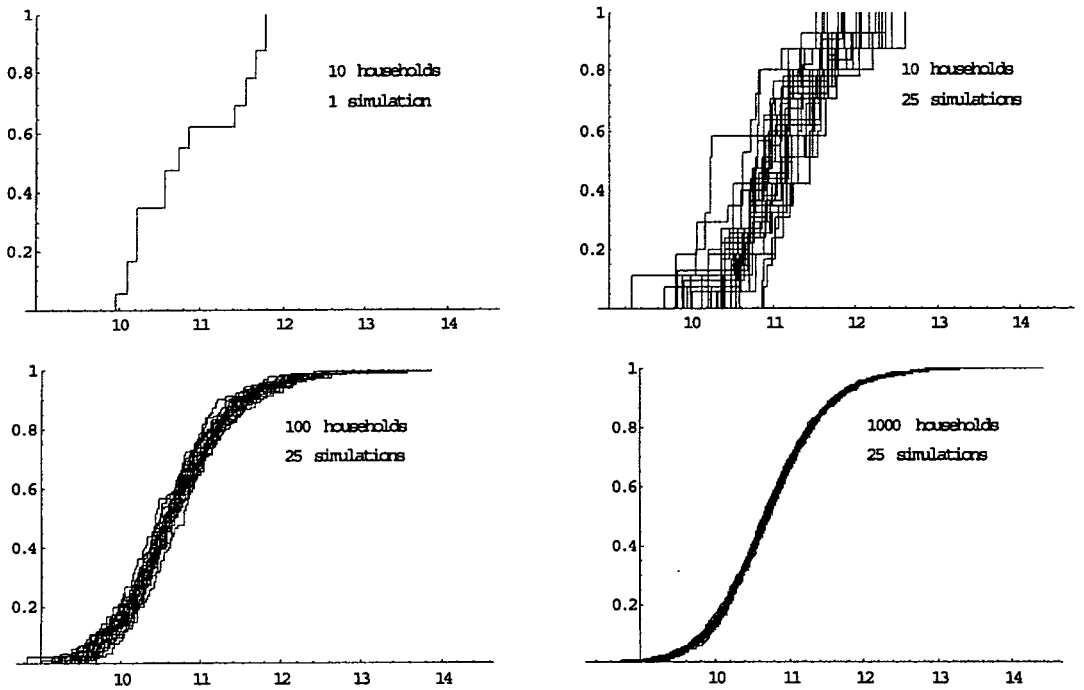
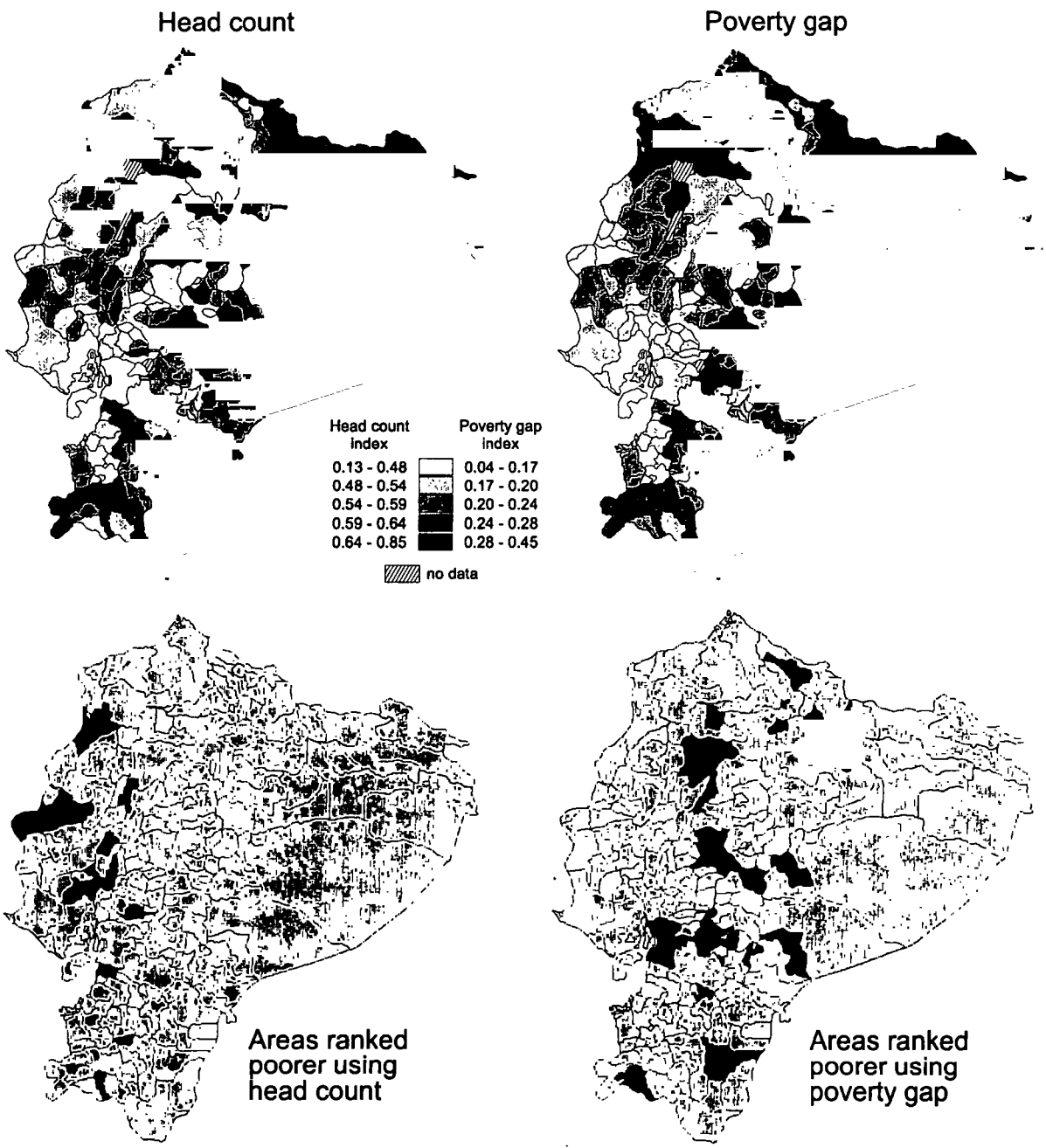


Figure 2



Idiosyncratic error falling with number of households in target population.

Figure 3[a]
Rural Poverty by Canton: Headcount and Poverty Gap



Note:

[a] The top two maps illustrate the geographical distribution of rural poverty across cantons based on respectively, the headcount measure of poverty and the poverty gap index. The shaded regions in the bottom two maps highlight those cantons where the rankings in the top two maps are not the same. The map on the left highlights those cantons that are ranked lower (more poor), according to the headcount measure, than they would be according to the poverty gap index. The map on the right highlights those cantons that are ranked lower according to the poverty gap index, than they would be according to the headcount measure.

Table 1: Diagnostics for Selected First-Stage Model Specifications

	Model			
Diagnostic	**I (Sparse) No Infrastructure No Means**	**II No Infrastructure Location Means**	**III Infrastructure No Means**	**IV (Full) Infrastructure Location Means**
Hausman test of Population weights (Deaton, 1997) H_0: $\beta^W = \beta^{NW}$	F-test: 1.66 95% Critical value (18, 448)=1.42	F-test: 2.05 95% Critical value (23, 438)=1.53	F-test: 1.57 95% Critical value (21, 442)=1.57	F-test: 1.84 95% Critical value (26, 432)=1.50
$\overline{R}^2$	0.41	0.47	0.42	0.50
Importance of random effect $\hat{\sigma}_\eta^2 / \hat{\sigma}_u^2$	0.141	0.048	0.149	0.019
H_0: Location effects jointly = 0 **p-value**	<0.001	0.024	<0.001	0.235
H_0: $\sigma_{\varepsilon,ch}^2 = \sigma_\varepsilon^2$ **p-value**	0.98	< 0.001	<0.001	< 0.001
Distribution of: $\hat{\eta}_c$				
Skewness	0.52	0.12	0.38	0.25
Kurtosis	3.06	0.87	3.91	0.35
$\chi^2(2)$ **- test of normal distribution**	1.79	7.47	2.27	11.85
Distribution of: $\hat{\varepsilon}_{ch}$	(N=483)	(N=484)	(N=484)	(N=484)
Skewness	-0.22	-0.48	-0.14	-0.51
Kurtosis	5.67	6.23	7.14	3.79
$\chi^2(2)$ **- test of normal distribution**	147.44	229.90	346.49	33.51

Table 2: Headcount and General Entropy (0.5) Measures – Different Consumption Models[a]

Model	Estimates	Headcount	GE (0.5)
I **Full Model** **Location Effect at EA Level**	No. Draws R	300	300
	$\hat{\mu}$	**0.508**	**0.275**
	Estimated Standard Error	**0.024**	**0.020**
	Due to[b]: Model	0.023	0.004
	Idiosyncratic	0.005	0.019
II **Full Model** **Location Effect at *Parroquia* Level**	**$\hat{\mu}$**	**0.508**	**0.504**
	Estimated Standard Error	**0.025**	**0.040**
	Due to: Model	0.021	0.004
	Idiosyncratic Error	0.013	0.039
III **Sparse Model** **Location Effect at EA Level**	**$\hat{\mu}$**	**0.504**	**0.514**
	Estimated Standard Error	**0.029**	**0.038**
	Due to: Model	0.028	0.008
	Idiosyncratic Error	0.009	0.037
IV **Sparse Model** **Assumption of No Location Effect**	**$\hat{\mu}$**	**0.526**	**0.521**
	Estimated Standard Error	**0.021**	**0.016**
	Due to: Model	0.020	0.007
	Idiosyncratic Error	0.004	0.015

Notes:

[a] These are household groups drawn randomly from the rural Costa census population as described in the text. The 'population' samples are of 15,000 households.

[b] These are the estimated standard deviations for each separate piece of the total variance, V_M and V_I.

Table 3: Headcount and General Entropy (0.5) Measures – Different Population Sizes

Model	Estimates	Number of households[a]			
		100	1,000	15,000	100,000
Headcount	$\hat{\mu}$	0.46	0.50	0.51	0.51
	Total Standard Error	0.067	0.039	0.024	0.024
	VI / Total Variance	0.75	0.24	0.04	0.02
GE (0.5)	$\hat{\mu}$	0.26	0.28	0.28	0.28
	Total Standard Error	0.048	0.029	0.022	0.022
	VI / Total Variance	0.79	0.28	0.03	<0.01

Notes:

[a] These are household groups drawn randomly from the rural Costa census population as described in the text. Smaller 'population' samples are subsets of the larger 'populations'.

[b] These are the estimated standard deviations for each separate piece of the total variance, V_M and V_I.

Table 4: Further Diagnostics

Comparison of *Parroquia*-level Estimates[a,b]	Estimation assuming homoscedasticity of disturbance components and no location effect	Estimation without use of population expansion factors	Estimation with semi-parametric disturbance distribution
Headcount			
Spearman's rank correlation $(\mu^*, \hat{\mu})$	0.98	0.73	0.998
$(\mu^* - \hat{\mu})$ **Mean absolute difference**	0.017	0.053	0.008
Minimum	-0.069	-0.169	-0.024
Maximum	0.023	0.306	0.026
General Entropy (0.5)			
Spearman's rank correlation $(\mu^*, \hat{\mu})$	0.86	0.91	0.957
$(\mu^* - \hat{\mu})$ **Mean absolute difference**	0.016	0.026	0.011
Minimum	-0.174	-0.151	-0.047
Maximum	0.067	0.009	0.192

Notes:

[a] $\hat{\mu}$ are estimates using the full model in column 4 of Table 1. μ^* are estimates which differ as indicated in the column headings.

[b] Comparisons are made of 271 *parroquias* in the rural Costa region.

Table 5: Improvement using Combined Data - Headcount

Region	Sample Data Only (region)		Combined Data (sub-region)	
	(2)	(3)	(4)	(5)
	S.E. of Estimate	Population (1000s)	S.E. of Estimate (median)	Population (median, 1000s)
Rural Sierra	0.027	**2,509**	0.038	**3.3**
Rural Costa	0.042	**1,985**	0.046	**4.6**
Rural Oriente	0.054	**298**	0.043	**1.2**
Urban Sierra	0.026	**1,139**	0.026	**10.0**
Urban Costa	0.030	**1,895**	0.031	**11.0**
Urban Oriente	0.050	**55**	0.027	**8.0**
Quito	0.033	**1,193**	0.048	**5.8**
Guayaquil	0.027	**1,718**	0.039	**6.5**

Table 6: Other Measures of Welfare

Measure	Estimates	Number of households[a]			
		100	1,000	15,000	100,000
FGT (1) Poverty Gap	$\hat{\mu}$	**0.159**	**0.176**	**0.176**	**0.176**
	Estimated Standard Error	**0.030**	**0.016**	**0.013**	**0.013**
	Due to[b]: Model	0.013	0.013	0.012	0.012
	Idiosyncratic	0.026	0.010	0.002	0.002
Variance of Log Per-capita Expenditure	$\hat{\mu}$	**0.453**	**0.480**	**0.480**	**0.482**
	Estimated Standard Error	**0.071**	**0.044**	**0.037**	**0.037**
	Due to: Model	0.037	0.039	0.037	0.037
	Idiosyncratic Error	0.060	0.021	0.006	0.002
Atkinson Index (2)	$\hat{\mu}$	**0.368**	**0.389**	**0.390**	**0.391**
	Estimated Standard Error	**0.046**	**0.028**	**0.024**	**0.023**
	Due to: Model	0.024	0.024	0.024	0.023
	Idiosyncratic Error	0.039	0.014	0.004	0.001

Notes:
[a] [b] See notes to Table 3.

Table 7
Decomposition of Inequality in Rural Ecuador by Regional Sub-Group
General Entropy (0.5)

Level of Decomposition	No. of sub-groups	Within-Group (%)	Between-Group (%)
National	1	100.0	0
Sector and: Region (Costa, Sierra, Oriente)	3	100.0	0
Province	21	98.7	1.3
Canton	195	94.1	5.9
Parroquia	**915**	**85.9**	**14.1**
Household	960,529	0	100.0

Appendix 1: The Estimator $\widehat{\sigma}_\eta^2$ and its Distribution

Estimation using moment conditions

For $c = 1, \ldots, C; h = 1, \ldots, n_c$, let η_c and ε_{ch} be independent random variables with zero expectation and finite variance, where the η_c are identically distributed. Suppose we have observations on u_{ch}, where

(25) $u_{ch} = \eta_c + \varepsilon_{ch}$.

The problem is to estimate $\sigma_\eta^2 = \text{var}(\eta)$. Using '.' to indicate the arithmetic mean over an index, (e.g., $\varepsilon_{c.} = 1/n_c \sum_h \varepsilon_{ch}$) we note that

$$u_{c.} = \eta_c + \varepsilon_{c.},$$

Hence

(26) $\text{E}[u_{c.}^2] = \sigma_\eta^2 + \text{var}(\varepsilon_{c.}) = \sigma_\eta^2 + \tau_c^2$.

We use the following lemma:

Lemma 1 *For $i = 1, \ldots, n$, let x_i be independent random variables with zero mean and finite variance, and let $\lambda_1, \ldots, \lambda_n$ be a given set of non-negative numbers, satisfying $\sum_{i=1}^n \lambda_i = 1$. Let $x_. = \sum_i \lambda_i x_i$ be the weighted average of the x_i. Then*

$$E[\sum_i \lambda_i (x_i - x_.)^2] = \sum_i \lambda_i (1 - \lambda_i) E[x_i^2].$$

The lemma implies that, for a set of non-negative weights w_c, summing to 1:

$$(27)\quad \mathrm{E}[\sum_c w_c(u_{c.} - u_{..})^2] = \sum_c w_c(1-w_c)(\sigma_\eta^2 + \tau_c^2).$$

Hence:

$$(28)\quad \sigma_\eta^2 = \frac{\mathrm{E}[\sum_c w_c(u_{c.} - u_{..})^2]}{\sum_j w_j(1-w_j)} - \frac{\sum_c w_c(1-w_c)\tau_c^2}{\sum_j w_j(1-w_j)}.$$

Note that

$$(29)\quad \tau_c^2 = \mathrm{var}(\varepsilon_{c.}) = \mathrm{E}[\frac{1}{n_c(n_c-1)}\sum_h(\varepsilon_{ch} - \varepsilon_{c.})^2].$$

A natural candidate for an estimator for σ_η^2 is therefore

$$(30)\quad \widehat{\sigma}_\eta^2 = \max(\frac{\sum_c w_c(u_{c.} - u_{..})^2}{\sum_j w_j(1-w_j)} - \frac{\sum_c w_c(1-w_c)\widehat{\tau_c^2}}{\sum_j w_j(1-w_j)}; 0),$$

where

$$(31)\quad \widehat{\tau_c^2} = \frac{1}{n_c(n_c-1)}\sum_h(\varepsilon_{ch} - \varepsilon_{c.})^2.$$

An estimator for the variance of $\widehat{\sigma}_\eta^2$ can be obtained using simulation (see below). As an alternatively, to approximate var $\left(\widehat{\sigma}_\eta^2\right)$ we make the following simplifying assumptions:

- $\varepsilon_{ch} \sim \mathcal{N}\left(0, \sigma_{\varepsilon,c}^2\right)$, homoskedastic within cluster.
- $\eta_c \sim \mathcal{N}\left(0, \sigma_\eta^2\right)$
- $u_{c.}^2$ and $\widehat{\tau_c^2}$ treated as independent and

- $u_{..} = 0$.

Denote $a_c = w_c / \sum_j w_j(1-w_j)$, $b_c = w_c(1-w_c)/\sum_j w_j(1-w_j)$, then $\widehat{\sigma}_\eta^2 = \sum_c a_c u_{c.}^2 - \sum_c b_c \widehat{\tau_c^2}$.

$$(32)\quad \text{var}\left(u_{c.}^2\right) = \text{var}\left(\eta_c^2 + \varepsilon_{c.}^2 + 2\eta_c\varepsilon_{c.}\right) = \text{var}(\eta^2) + \text{var}(\varepsilon_{c.}^2) + 4\sigma_\eta^2\tau_c^2.$$

Note that under the assumptions above, $\widehat{\tau_c^2}$ is distributed as $\tau_c^2\chi_{n_c-1}^2/(n_c-1)$, hence its variance is

$$(33)\quad \text{var}(\widehat{\tau_c^2}) = 2\frac{\tau_c^4}{n_c-1}.$$

Similarly, $\varepsilon_{c.}^2$ is distributed as $\tau_c^2\chi_1^2$ with variance $2\tau_c^4$ and $\text{var}(\eta^2) = 2\sigma_\eta^4$.

Combining, we find

$$(34)\quad \text{var}\left(\widehat{\sigma}_\eta^2\right) \approx \sum_c [a_c^2 \text{var}\left(u_{c.}^2\right) + b_c^2 \text{var}(\widehat{\tau_c^2})] \approx \sum_c 2[a_c^2\{(\widehat{\sigma}_\eta^2)^2 + (\widehat{\tau_c^2})^2 + 2\widehat{\sigma}_\eta^2\widehat{\tau_c^2}\} + b_c^2\frac{(\widehat{\tau_c^2})^2}{n_c-1}].$$

Estimation using simulation

The following more direct approach can also be taken.

- Estimate σ_η^2 from equation (30) above. This gives $\widehat{\sigma_\eta^2}$.
- Estimate $\sigma_{\varepsilon,ch}^2$ heteroskedasticity model in Section 3. This gives $\widehat{\sigma}_{\varepsilon,ch}^2$.

- Using the estimated variance components, and assuming η_c and ε_{ch} to be independent and normally distributed with mean zero, generate new values for u_{ch}, using equation (25).

- Compute a new estimate for σ_η^2 using formula (30).

- Repeat many times, keeping the simulated values of σ_η^2.

The set of simulated values for σ_η^2 thus obtained can be used to calculate the sampling variance of $\widehat{\sigma_\eta^2}$ directly.

In practice $\widehat{\sigma_\eta^2}$ is often so small that equation (30) will generate a significant number of zero variance estimates for η (i.e., the distribution is far from normal). Given this feature of the sampling distribution of $\widehat{\sigma}_\eta^2$, using *only* information on the point estimate and its sampling variance could be misleading (as when using the delta method to calculate the model variance, V_M). The alternative approach to calculating the variance of $\widetilde{\mu}$ discussed following equation (16) could be implemented by taking random draws of σ_η^2 from the set of simulated values of σ_η^2 obtained above,therefore using the full distribution.

Appendix 2: First Stage Regression Results

Table A.1.

First-Stage Estimates for Log Per-Capita Expenditure: Rural Costa

Variable[a]	Parameter estimate[b]	Estimated Standard Errors
I. Household-level/ Non-Infrastructure		
Family size	-0.623	0.0947
Family size squared	0.062	0.0138
Family size cubed	-0.002	0.0006
Indigenous language spoken	0.004	0.0035
Rented home	0.001	0.0015
Owned home	0.002	0.0005
Walls of brick	0.002	0.0007
Walls of wood	-0.002	0.0008
Cooking on gas fire	0.0001	0.0019
Cooking with wood or charcoal	-0.0008	0.0019
Persons per bedroom	0.049	0.1018
Persons per bedroom squared	-0.014	0.0185
Persons per bedroom cubed	0.0007	0.0009
Household head with no spouse	-0.089	0.1500
Years of schooling of:		
Household head	0.027	0.0067
Spouse of head	0.011	0.0084
Age of:		
Household head	0.005	0.0025
Spouse of head	-0.002	0.0030
II. Household-level/ Infrastructure		
Own connection to modern sewage	0.002	0.0005
Shared connection to modern sewage	0.0005	0.0010
Own latrine	0.0002	0.0006
III. Location Means/ Non-Infrastructure		
Age of household head	-0.026	0.0064
Years of schooling of spouse of head	-0.098	0.0327
% of household heads male	-0.025	0.0054
(Persons per bedroom)^2	0.019	0.0043
IV. Location Means/ Infrastructure		
Own connection to modern sewage	0.004	0.0012
Number of household observations	485	
Number of sample clusters	39	

Notes:

[a] Age and education for a child in a specific birth position is set equal to zero if the household does not have such a child. The location mean variables are household values of the indicated variable in the census data averaged over all households in a census enumeration area. ^2 indicates that the mean is squared. Dummy variables are defined as either 100 or 0.

[b] Parameters and standard errors are two-step GLS estimates calculated using household expansion factors and estimated variances of the disturbance components σ_η and σ_ε.

Table A.2
Model of Heteroscedasticity in ε_{ch}[a]

Variable	Parameter Estimate	Estimated Standard Errors
Constant	-4.161	0.427
Years schooling of head's spouse	-2.516	1.066
Wood walls	0.018	0.004
Predicted log per capita expenditure * spouse education	0.299	0.083
Head's education * age of head	-0.005	0.002
Head's education * cooking with gas	0.001	0.0007
Age of head * education of spouse	0.019	0.009
Spouse's education * age of spouse	-0.009	0.003
Spouse's education * crowding	-0.525	0.150
Spouse's education * own latrine	0.001	0.0006
Age of Spouse ^ 2	0.0004	0.0001
Shared sewage connection * brick walls	-0.0002	0.00005
Head with no spouse * rented home	0.044	0.004
Spouse's education * household size	0.059	0.018
Spouse's education * (crowding^2)	0.104	0.029
Spouse's education * (crowding^3)	-0.006	0.002
Own sewage connection * (crowding^3)	-0.00003	0.00003
Brick walls * (household size^3)	0.00004	0.00001
Wooden walls * (crowding^3)	-0.00008	0.00002
Gas cooking * (household size^3)	-0.00004	0.00001
Gas cooking * (crowding^3)	0.00004	0.00001
$\overline{R}^2$	0.25	

Note:

[a] The dependent variable is $(\hat{u}_{ch} - \hat{u}_{c.})^2$. See notes to Table A.1 for other variable definitions. The model and standard errors are estimated using household expansion factors. Standard errors are White robust estimates.

Policy Research Working Paper Series

	Title	Author	Date	Contact for paper
WPS2895	Telecommunications Reform in Côte d'Ivoire	Jean-Jacques Laffont Tchétché N'Guessan	September 2002	P. Sintim-Aboagye 38526
WPS2896	The Wage Labor Market and Inequality in Vietnam in the 1990s	John Luke Gallup	September 2002	E. Khine 37471
WPS2897	Gender Dimensions of Child Labor and Street Children in Brazil	Emily Gustafsson-Wright Hnin Hnin Pyne	October 2002	M. Correia 39394
WPS2898	Relative Returns to Policy Reform: Evidence from Controlled Cross-Country Regressions	Alexandre Samy de Castro Ian Goldin Luiz A. Pereira da Silva	October 2002	R. Yazigi 37176
WPS2899	The Political Economy of Fiscal Policy and Economic Management in Oil-Exporting Countries	Benn Eifert Alan Gelb Nils Borje Tallroth	October 2002	J. Schwartz 32250
WPS2900	Economic Structure, Productivity, and Infrastructure Quality in Southern Mexico	Uwe Deichmann Marianne Fay Jun Koo Somik V. Lall	October 2002	Y. D'Souza 31449
WPS2901	Decentralized Creditor-Led Corporate Restructuring: Cross-Country Experience	Marinela E. Dado Daniela Klingebiel	October 2002	R. Vo 33722
WPS2902	Aid, Policy, and Growth in Post-Conflict Societies	Paul Collier Anke Hoeffler	October 2002	A. Kitson-Walters 33712
WPS2903	Financial Globalization: Unequal Blessings	Augusto de la Torre Eduardo Levy Yeyati Sergio L. Schmukler	October 2002	P Soto 37892
WPS2904	Law and Finance: Why Does Legal Origin Matter?	Thorsten Beck Asli Demirgüç-Kunt Ross Levine	October 2002	K. Labrie 31001
WPS2905	Financing Patterns Around the World: The Role of Institutions	Thorsten Beck Asli Demirgüç-Kunt Vojislav Maksimovic	October 2002	K. Labrie 31001
WPS2906	Macroeconomic Effects of Private Sector Participation in Latin America's Infrastructure	Lourdes Trujillo Noelia Martín Antonio Estache Javier Campos	October 2002	G. Chenet-Smith 36370
WPS2907	The Case for International Coordination of Electricity Regulation: Evidence from the Measurement of Efficiency in South America	Antonio Estache Martin A. Rossi Christian A. Ruzzier	October 2002	G. Chenet-Smith 36370
WPS2908	The Africa Growth and Opportunity Act and its Rules of Origin: Generosity Undermined?	Aaditya Mattoo Devesh Roy Arvind Subramanian	October 2002	P. Flewitt 32724
WPS2909	An Assessment of Telecommunications Reform in Developing Countries	Carsten Fink Aaditya Mattoo Randeep Rathindran	October 2002	P. Flewitt 32724

Policy Research Working Paper Series

	Title	Author	Date	Contact for paper
WPS2910	Boondoggles and Expropriation: Rent-Seeking and Policy Distortion when Property Rights are Insecure	Philip Keefer Stephen Knack	October 2002	P. Sintim-Aboagye 38526