Chapter 2

Combining census and survey data to construct a poverty map of South Africa

Harold Alderman, Miriam Babita, Jean Lanjouw, Peter Lanjouw, Nthabiseng Makhatha, Amina Mohamed, Berk Özler and Olivia Qaba*

Introduction

Geographical dimensions of poverty inform both public policies on, and research into the determinants of, economic development and poverty. Poverty maps, for example, are used in many developing countries to allocate resources to local agencies or administrations as a first step in reaching the poor. Similarly, ranking of community needs is a step towards prioritising programmes. However, in practice, these measures have only been useful at fairly aggregated levels. The effectiveness of using locale as a means of directing resources to the poor is a function of the level of the geographic unit chosen for allocation. This works best when the unit is relatively small (Baker and Grosh, 1994).

Globally, information on many aspects of living standards, especially poverty measured by household income or expenditure, is rarely available for a sufficient number of households to permit construction of a finely disaggregated map, or for ranking local units of government based on poverty levels. For example, the World Bank's living standard measurement surveys (LSMS), variants of which have been fielded in many developing countries, do not allow for disaggregation of average incomes or of poverty rates much beyond a simple rural/urban breakdown within broad regions of a given country.

Unlike most sample surveys, census data do not suffer from small sample problems. However, they typically contain little direct information on household resources. The lack of income or expenditure information in such data sets has often prompted policy makers to explore alternative welfare indicators to derive the required geographic dimension of poverty and inequality. Many countries have developed sometimes crude, sometimes more sophisticated, basic needs indicators for this purpose but these indicators do not always conform well with consumption or income welfare indicators (Grosh and Glinskaya, 1997, Hentschel *et al.*, 1999).

In other countries, including South Africa as well as Australia, income classifications are obtained in the census by using broad ranges. The classification of individual or household income into such ranges seldom conveys to the respondent a clear definition of income. Thus, even abstracting from the nearly universal tendency of households to conceal income from interviewers, a respondent may fail to consider key components of income for typically poor households, such as agricultural profits (either from sale or own consumption) or informal sector profits and casual wages. Again, this measure of income may not be a fair indicator of income and consumption.

This motivates the interest in seeking ways to combine the detailed information obtained in household surveys with the more extensive coverage of a census to derive detailed geographic poverty estimates based on a consumption welfare indicator. This has recently been explored by

* The authors wish to thank Deon Filmer and Charles Simkins for helpful comments on an earlier draft, and especially Gabriel Demombynes for assistance with large portions of the analysis.

Hentschel *et al*. (2000) and Elbers *et al*. (2000), who both model consumption behaviour from a household survey in Ecuador, using a set of explanatory variables that are restricted to those also available in the Ecuadorian census. Applying the resulting parameter estimates to the census, both papers show how the probability that a given household in the census is in poverty can be derived. These authors also show how detailed geographic poverty rates can be calculated. Elbers *et al*. also provide a comprehensive description of the methodology they used in their study.

Information on aspects of living standards at a disaggregated level has a particular function in South Africa since the constitution requires parliament to pass legislation providing for the equitable division of nationally raised revenue among provincial and local spheres of governments. In terms of the Division of Revenue Act (Act 28 of 1998) passed in March 1998, provision is made for the distribution of a grant to municipalities – of which there were, at the time of writing, 843 – based on levels of poverty. This equitable shares grant is an unconditional grant to the municipality and is not a transfer to households intended to bring their incomes up to a target level. Nevertheless, the grant is based, in part, on the number of households within the jurisdiction which have an income of less that R800 per month.[1] However, there is no direct means of assessing the number of individuals in this category. This key allocation must be performed using incomplete or indirect information. As a general rule, central governments may not have the capacity to obtain this type of information directly and local governments may not have the incentive to transmit it (Alderman, 1999).

This study builds on the approach described above in order to utilise information from the 1995 South Africa October household survey (OHS) and the related income and expenditure survey (IES) in conjunction with the 1996 population census. We present evidence that incomes and poverty rates reported in the census differ systematically from those obtained in the household survey. We provide an alternative imputed expenditure estimate that is both consistent with the survey estimates and available for virtually all households which appear in the census. Thus, the methodology illustrates a means to obtain expected poverty estimates at any sub-national level of administration for which the information is desired.

The next section provides more details on the methodology and its links to the literature. In a further section relevant features of the data sets employed in this study are discussed. The section thereafter presents some direct comparisons between the mean levels of income and expenditure and poverty rates from the IES at various levels of aggregation and the corresponding means and poverty rates from Census '96. A subsequent section presents results of the regressions of consumption on housing and access to services, which form the basis for the imputation of consumption in the census data. The analogous comparisons to the third section are repeated using these imputations. In the next section the poverty mapping exercise is discussed. In a penultimate section the way forward in cooperative work between Stats SA and the World Bank are outlined. A final section draws the results together. The appendix provides the estimates of expected poverty rates, measured by the headcount index, and their standard errors, by province, by district council, and by magisterial district.

[1] Further information on this grant can be obtained from the South African local government website at: http:/www.local.gov.za/DCD/dcdindex.html

Methodology

The basic methodology applied in linking surveys and census-type data sets is very similar to that of synthetic estimation used in small-area geography. Prediction models are derived for consumption or income as the endogenous variable, on the basis of the survey. The selection of exogenous variables is restricted to those variables that can also be found in the census (or some other large data set). The parameter estimates are then applied to the census data and expected poverty and inequality statistics derived. Simple performance tests can be conducted which compare basic poverty or inequality statistics across the two data sets. For Ecuador, Hentschel *et al*. (2000) show that regional poverty estimates, calculated on the basis of imputed household consumption in the census, are very similar to those derived from consumption measured directly in the household survey.

The calculation of expected poverty and inequality statistics using predicted income or consumption has to take into account that each individual household income or consumption value has been predicted and has standard errors associated with it. Elbers *et al*. (2000) show that the approach yields estimates of the incidence of poverty and of inequality that are unbiased, and that the standard errors are small. Furthermore, the Ecuador case study demonstrates that these estimates are quite precise to permit meaningful comparisons across regions, and that the confidence intervals do not widen further with higher levels of spatial disaggregation provided that the population of the unit of disaggregation remains sufficiently large.[2]

The combination of information from different data sets has sparked a recent interest in the literature, e.g. Arellano and Meghir (1992), Angrist and Krueger (1992) and Lusardi (1996). Typically, however, these studies combine several household surveys rather than surveys with census data, and so far they have not been used to study spatial dimensions of poverty. While within-sample imputation of missing observations is a quite common procedure, e.g. Paulin and Ferraro (1994), out-of-sample imputation, which combines different data sets, is less frequent. One recent study that does combine an expenditure survey with census information to estimate local income distributions is Bramley and Smart (1996). However, this study differs from the approach used here in that Bramley and Smart did not have access to unit level data from both data sources and hence derived local income distributions not from predicted household incomes but from estimates of mean incomes of different locale and distribution characteristics.

This study differs from other studies in the literature, including Hentschel *et al*. (2000) in that, while we are imputing values for consumption which are not present in the census, we are also substituting them for a variable, income, for which estimates are available. By what measure do we know we have substituted an improved indicator of the welfare of the community? We will take as a maintained hypothesis that consumption is generally more accurately collected in household surveys than is income and that it is a valid measure of the long run control of resources by the household (Deaton,

[2] Hentschel *et al*. (1999) state that: 'In fact, a poverty map would have to be constructed at quite a high degree of spatial disaggregation before the standard errors increase significantly due to small populations … Only when the [local] population falls well below 500 households does the corresponding standard error rise to levels which could compromise comparisons.'

1997).[3] Thus, we seek to compare the correspondence of both the average of the income measure obtained in the census and the poverty rates calculated using this measure with those estimates using the expenditure measure in the IES. If the imputation of expenditure is of value then the imputed measure using census data should be closer to the IES indicators of consumption and poverty. In addition to looking at the correlation of poverty measures and rankings on poverty we also look at a measure of the fit based on the absolute difference between the two poverty measures. This is defined as

$$\text{Fit} = 1/N[\Sigma \left| Y_i - \hat{Y}_i \right| / \text{mean}(Y_i)]$$

where Y_i is a measure of poverty derived using IES data (poverty rate, average expenditures, or income) for a given unit, denoted by the subscript i. Similarly, $\hat{Y}$ indicates the corresponding estimate from the census.

While the goodness of fit measure provides a summary statistic, we also regress the individual components of the statistic against variables that may account for differences in the accuracy of the census income data. That is, we run regressions using $\left| Y_i - \hat{Y}_i \right| / \text{mean}(Y_i)$ as the left hand variable. This allows us to investigate whether the bias in average reported census income, measured by its divergence from mean expenditure in the household survey for the same region, varies between areas depending, among other factors, on the sectoral composition in each region.

The levels of administrative units in South Africa, in order of higher disaggregation, are as follows: province, district council, magisterial district, and urban or rural place name. At the time of writing, there were nine provinces, 45 district councils, 354 magisterial districts (MDs), and 12 753 towns or place names. The validation, however, must take into account that the IES was not designed to be representative at levels of disaggregation for which we want to use the data. Indeed, were it representative for lower levels of administration there would be little need to impute expected poverty estimates into the census. Thus, although we can link the OHS and the census at the magisterial district level, validation using this imprecise, albeit unbiased, reference point is of limited value. For this reason, we first perform our validation exercise at the province level even though we seek to create a poverty map for smaller geographical units. We repeat the exercise, however, at higher degrees of spatial disaggregation mainly to demonstrate what happens to the goodness of fit measure at lower levels of administration. Hence, we calculate mean census income and mean imputed expenditure in the census for each province and determine how they fare against the mean household expenditure in the IES for the corresponding province.

[3] We focus on the best means of measuring income or consumption poverty and abstract from the debate those measures of household welfare which add to a multi-dimensional understanding of poverty. See Ravallion (1992) for further discussion on the measurement of poverty.

Data

This section provides some information on each of the three data sources that are utilised.

The OHS is an annual survey, which focuses on a few key indicators of living patterns in South Africa. In particular the survey focuses on employment, internal migration, housing, access to services, individual education, and vital statistics. In the 1995 round of the survey, 29 700 households were interviewed.

As its name implies, the IES provides information on the income and expenditure of households for the 12-month period prior to the interview. The questionnaire was designed to capture the value of gifts and in-kind benefits and the imputed value of housing under income and consumption. The following information provides some ideas about the detail of consumption data collected. The cost of housing is based on 27 questions and monthly expenditures on food and beverage is aggregated up from information obtained in 131 questions. Twenty-two additional questions cover food consumed from own production. Similar details are sought regarding non-food purchases and services obtained, using a mix of monthly and annual recall. The expenditure variable used in this study is slightly redefined from standard Stats SA reporting from the 1995 IES. In order to correspond more closely to current consumption as a standard measure of household welfare, we netted out income taxes as well as various forms of saving (including lumpy purchases of durable goods and vehicles as well as *lobola* and dowry) from the total expenditures.

Income is based both on individual formal and non-formal earnings and returns to household assets as well as gifts and dowry received. In order to make these income and consumption aggregates comparable with the census data, all incomes and expenditures were put into 1996 Rand using the consumer price index.

The IES was designed to be merged with the OHS. While the interviews for the IES were conducted at a slightly later date than the OHS, the same households were visited. In all, 28 585 households remained in the data set after the two surveys were merged.

Census '96 covers over nine million households, recording data from individuals based on where they were the night between 9 and 10 October 1996. In addition to information on household composition, it collected some details on housing and services in a manner that paralleled the OHS. It also asked every *individual* to indicate his or her income, including pensions and disability grants. The individuals were asked to indicate which of 14 brackets this income fell within. In order to get to household income, each of these ranges was assigned a point value. For most categories this value was the logarithmic mean of the top and bottom income of the bracket. For the lowest group with income, however, the value was two-thirds of the interval. For the highest bracket (greater than R360 000 per year) this value was 720 000. These assignments follow standard practice within Statistics South Africa. The census also asks for the value of all remittances received by the household in the preceding year. The individual point estimates for each bracket were then summed. This figure was added to the estimate of household income.

All of these data sets include coding for the province, the enumeration area type (EA type), the district council, and the magisterial district in which the household resided. These geographic units are the ss

units of analysis in this study. As mentioned above, only the provinces are representative of the sample, but given how the sample was stratified, the breakdown to EA type within each province should also be quite close to being representative of the breakdown of the population into residents of urban portion of former homelands, other rural residents, urban formal, urban informal and other types of enumeration areas.[4] At each level of disaggregation, we excluded from our analysis units where three or less enumeration areas were revisited in the household survey.

For both the IES and Census '96 we averaged income per household and per capita over each of our units of analysis.[5] We also created headcount poverty indices for each geographical unit. This index is the well-known Foster, Greer and Thorbeck poverty measure (FGT) defined as

$$P_{i_\alpha} = \frac{1}{N}\sum_{h=1}^{N}\left(\frac{z - y_h}{z}\right)^{\alpha} \mid (y_h \leq z)$$

where P_i is the index of poverty for the ith magisterial district, y_h is a measure of household income from a sample of size N and z is the poverty line. With the headcount index α is zero, while it is set to one to measure poverty gap and higher for the severity of poverty. While this study focuses on the headcount measure of poverty, the methodology can be applied to these measures as well. The FGT measure is additive. Thus, one can go from poverty in each magisterial district to a consistent indicator of provincial or national poverty.

Comparing Census '96 income and IES expenditure

The average income from the IES is R3 309 per household per month, while the average monthly current expenditure is R2 954.[6] Both these estimates exceed the monthly income including remittances from the census income data. That average is R2 454. The IES *expenditure* figure aggregates up very close to the R330 billion of private consumption for 1996 estimated by the South African Reserve Bank, while the latter is nearly 20% below. In principal, household income includes private investment and, therefore, should exceed private consumption. Thus, the IES figures are fairly consistent with the share of gross national product (GNP) not accounted for by government consumption, corporate savings, or account deficits, while the aggregation from Census '96 is less so. Given the difference in income in the two data sets, it is not surprising that poverty rates using the IES also differ from those based on census data. We indicate this using two different poverty lines. One is the R800 per household per month line at which households are defined as poor for the purpose of the equitable shares grant. The second is a measure of per capita income set at R250. Using these two poverty lines and the expenditure data from the IES, the percentage of poor in the country is 28,4 and

[4] The sample was stratified by province, urban and non-urban areas, and population group.

[5] Recent studies have indicated that the poverty ranking of households is sensitive to assumptions regarding the degree that households have scale economies as well as whether adult equivalency scales are assumed for children (Lanjouw, Milanovic and Paternostro, 1999). However, we do not address this possibility in the current study.

[6] These averages were calculated using sampling weights that were available at the province level. For averages that were calculated for administrative units smaller than a province, such as district councils or magisterial districts, no sampling weights were used because they were not available.

48,4 respectively.[7] However, using the income from the census, the estimated number of poor based on the *household poverty line* is 52,2%. That is, the estimated poverty rate is over 80% higher in the census than the IES data. Similarly, using the *per capita poverty line*, the poverty rate from the census at 60,8% is also larger than that estimated from the IES.

The difference between the census and IES poverty estimates reported above can not be attributed to the fact that the former are based on incomes while the latter are based on expenditures. Poverty estimates using the *income* data from the IES show the percentage of poor in the country are 28,6 and 46,2 for the two poverty lines. Thus, the estimated rates of poverty are very similar to those estimated using expenditures. Given the close correspondence of the poverty estimates using either income of expenditure based on IES data, we will for the remainder of this paper concentrate on the expenditure data from the IES.

As indicated in Table 1, six out of the nine province-level income averages from the IES are significantly different to their counterparts from the census. However, this does not necessarily mean a poor correlation of average incomes by province as defined in the census with the average expenditures by province from the IES. While the correlation coefficient between the census income and IES expenditure is 0,93, the ordering in terms of income differ, hence the Spearman rank correlation coefficient is only 0,68 (see Table 2). The corresponding figures for the poverty measures in terms of the percentage of households with less than R800 per month calculated from the two alternative data sources are 0,76 and 0,55, respectively. While there is still a large difference in provincial poverty rates between the census and the IES when using the per capita poverty expenditure line of R250 per capita, the correlation coefficient rises to 0,93 although the rank correlation coefficient is only 0,72.

Table 1: Comparison of household income from Census '96 and household expenditure from the IES

Province	Mean hh income (Rand/month) [census]	Mean hhs exp. (Rand/month) [IES]	% of hhs with monthly income below R800 [census]	% of hh with monthly exp. below R800 [IES]	% of individuals in hhs with per capita monthly income below R250 [census]	% of individuals in hhs with per capita monthly exp. below R250 [IES]
Western Cape	3 976	3 919 (181,40)	26,74*	12,45 (1,12)	30,09*	25,32 (1,80)
Eastern Cape	1 479*	1 815 (80,92)	68,30*	44,51 (1,40)	76,41*	67,93 (1,34)
Northern Cape	2 244	2 217 (164,90)	50,33*	38,02 (3,00)	59,11*	52,57 (2,96)
Free State	1 823	1 794 (106,30)	58,81*	51,04 (2,22)	66,25	62,16 (2,13)
KwaZulu-Natal	2 193*	2 680 (111,00)	55,37*	24,27 (1,36)	66,12*	52,17 (1,77)
North West	1 737*	2 218 (176,00)	56,06*	37,18 (2,40)	65,40*	58,88 (2,22)
Gauteng	4 044*	5 086 (221,50)	33,90*	10,57 (1,17)	34,34*	14,37 (1,43)
Mpumalanga	1 762*	2 356 (144,60)	60,19*	25,58 (2,17)	68,42*	53,96 (2,19)
Northern Prov.	1 234*	2 188 (130,90)	71,76*	36,42 (2,10)	79,93*	58,01 (2,17)

Standard errors in parentheses.
***Signifies statistically significant differences from census averages at the 5% level.**

[7] Note that the first figure is household poverty, while the latter is individual poverty, i.e. 28,8% of the households in South Africa have a monthly household income of less than R800, whereas 48,4% of the individuals live in households with monthly per capita income of less than R250.

Table 1A: Comparison of ***imputed*** expenditure from Census '96 and household expenditure from the IES

Province	Mean imputed hh expenditure (Rand/ month) [census]	Mean hh expenditure (Rand/month) [IES]	% of hhs with imputed monthly expenditure below R800 [census]	% of hhs with monthly expenditure below R800 [IES]	% of individuals in hhs with per capita monthly imputed expenditure below R250 [census]	% of individuals in hhs with per capita monthly expenditure below R250 [IES]
Western Cape	3 835	3 919 (181,4)	12,05	12,45 (1,12)	22,67	25,32 (1,80)
Eastern Cape	1 718	1 815 (80,92)	47,29	44,51 (1,40)	66,56	67,93 (1,34)
Northern Cape	2 400	2 217 (164,9)	35,04	38,02 (3,00)	49,78	52,57 (2,96)
Free State	1 795	1 794 (106,3)	48,14	51,04 (2,22)	60,47	62,16 (2,13)
KwaZulu-Natal	2 586	2 680 (111,0)	25,67	24,27 (1,36)	50,41	52,17 (1,77)
North West	2 188	2 218 (176,0)	37,32	37,18 (2,40)	52,76*	58,88 (2,22)
Gauteng	4 341*	5 086 (221,5)	13,20*	10,57 (1,17)	18,92*	14,37 (1,43)
Mpumalanga	2 391	2 356 (144,6)	24,46	25,58 (2,17)	46,33*	53,96 (2,19)
Northern Prov.	1 837*	2 188 (130,9)	37,44	36,42 (2,10)	59,93	58,01 (2,17)

Standard errors in parentheses.
***Signifies statistically significant differences from census averages at the 5% level.**

Table 2: Simple and rank correlation coefficients between Census '96 income and IES expenditure

	Number of observations	Simple correlation coefficient	Rank correlation coefficient	Correlation coefficient for poverty measures (hh poverty with z = R800)	Rank correlation coefficient for poverty measures (hh poverty with z = R800)
Provinces (census and IES)	9	0,9275 (0,0003)*	0,6833 (0,0424)*	0,7612 (0,0172)*	0,5500 (0,1250)
Provinces (imputed census and IES)	9	0,9790 (0,0000)*	0,9333 (0,0002)*	0,9887 (0,0000)*	0,9000 (0,0009)*
Province/EA type (census and IES)	31	0,9339 (0,0000)	0,7786 (0,0000)	0,6971 (0,0000)	0,6065 (0,0003)
Province/EA type (imputed census and IES)	31	0,9475 (0,0000)	0,8766 (0,0000)	0,8546 (0,0000)	0,8863 (0,0000)
District council (census and IES)	45	0,8844 (0,0000)	0,7835 (0,0000)	0,7145 (0,0000)	0,6872 (0,0000)
District council (imputed census and IES)	45	0,8844 (0,0000)	0,8407 (0,0000)	0,8603 (0,0000)	0,8672 (0,0000)
Magisterial district (census and IES)	354	0,7084 (0,0000)	0,6352 (0,0000)	0,5753 (0,0000)	0,5325 (0,0000)
Magisterial district (imputed census and IES)	354	0,6949 (0,0000)	0,6694 (0,0000)	0,6957 (0,0000)	0,7047 (0,0000)

Standard errors in parentheses.
*** denotes significance at the 5% level**

Census '96 collects income information from one question on individual income including pensions and one on remittances without any probing about informal income or enterprise profits. In contrast, the household survey details both income and expenditure information as described in the beginning of this section. As a result, the census income is understated for most of the population, but likely more in rural areas. That is, it is plausible that people in urban areas, with a higher share of individuals earning salaries, are able to state their earnings better than people who live in rural portions of former homelands or other rural areas, who earn more from casual income and from own production, according to Census '96.

This is explored with the regressions reported in the first four columns of Table 3 which demonstrate the fact that the gap between the IES and the census differs depending, among other things, on the urban/rural composition of the province.[8] All of these regressions have considerable explanatory power, measured by the adjusted R^2. This indicates that the measure of goodness of fit is correlated with other observable characteristics and that the gap between census income and IES expenditure varies by some of these characteristics. However, there are only nine provinces in these regressions. Therefore there is a problem regarding the degrees of freedom. Below we repeat these regressions at different levels of aggregation.

The first two columns in Table 3 show regression results for the goodness of fit of the estimate of average income at the province level defined above as a function of the percentage of population living in rural areas classified as former homelands (or as urban formal) as well as the average provincial expenditure using the IES data. The overall goodness of fit measure for the left-hand variable in the regression is 0,187, but ranges from 0,009 to 0,353 over the provinces. The larger the percentage of population residing in rural areas of former homelands in a province the less correspondence between the census and the IES data (i.e. the *higher* the figure for the goodness of fit) as indicated by the positive and statistically significant coefficient on the variable. Similarly, the coefficient on the variable for the urban formal areas is negative and significant.

Furthermore, controlling for area of residence, provinces with higher average expenditures also have a larger gap between census income and IES expenditure. Since we are dealing with only nine observations at this time, we can match this result with the data in Table 1. For example there is a large gap in Gauteng, despite the fact that 81% of its population lives in urban formal areas, which likely accounts for the coefficient on the variable for provincial average expenditure. For the two other provinces with no areas classified as former homelands (Western Cape and Northern Cape), there are no significant differences between the two measures. The goodness of fit measures for these two provinces are quite small being 0,019 and 0,009, respectively.

[8] We discuss the last four columns of Table 3, as well as Tables 4-6, after the methodology for imputing expenditures is presented.

Table 3: Regression of goodness of fit on area of residence and mean expenditure (province level)

Dependent variable: goodness of fit	Fit between census income and IES expenditure				Fit between imputed census exp. and IES expenditure			
	Mean expenditures		Headcount indices		Mean expenditures		Headcount indices	
	Coeff. (1)	Coeff. (2)	Coeff. (3)	Coeff. (4)	Coeff. (5)	Coeff. (6)	Coeff. (7)	Coeff. (8)
IES expenditure (,000)	0,088 (0,028)*	0,148 (0,028)**	0,132 (0,072)	0,309 (0,074)**	0,063 (0,021)*	0,074 (0,027)*	0,01 (0,015)	-0,2 (0,019)
% former homelands	0,414 (0,118)**		1,29 (0,306)**		0,098 (0,088)		-0,071 (0,062)	
% urban formal		-0,678 (0,134)**		-2,05 (0,355)**		-0,144 (0,131)		0,115 (0,091)
NF(2,6)	7,73	15,56	8,89	16,63	4,59	4,52	0,67	0,82
Adjusted R^2	0,627	0,784	0,664	0,796	0,473	0,468	-0,089	-0,048
N	9	9	9	9	9	9	9	9
Mean goodness of fit		0,183		0,849		0,081		0,061

Standard errors in parentheses.
*** denotes significance at the 5% level and**
**** at the 1% level.**

The difference between the census and IES poverty estimates reported above can not be attributed to the fact that the former are based on incomes while the latter are based on expenditures. Poverty estimates using the *income* data from the IES show the percentage of poor in the country are 28,6 and 46,2 for the two poverty lines. Thus, the estimated rates of poverty are very similar to those estimated using expenditures. Given the close correspondence of the poverty estimates using either income of expenditure based on IES data, we will for the remainder of this paper concentrate on the expenditure data from the IES.

The third and fourth columns of Table 3 show results of regressions using the goodness of fit of the head count of poverty. Again, the percentage of rural portions of former homelands is associated with a large gap between the census and the IES poverty estimates and the percentage of households in formal urban areas is associated with a better fit.

We repeat the analysis at higher levels of disaggregation, hence increasing the number of observations. First, we take the averages for income or expenditure and the poverty rates in each province separately if the enumeration area was defined as urban formal, urban informal, rural or former homeland. Since there are not former homelands in every province or a sufficient number of enumeration areas defined as 'urban informal', this provides 31 cells instead of the nine provincial averages. The regression in the first four columns of Table 4 indicate that the basic story is unchanged; the fit is less precise when the average is over a rural portion of former homeland and lower for urban formal. The goodness of fit also declines with a higher average expenditure.

Table 5 repeats these regressions with the unit of observation being the goodness of fit with income averaged over 45 district councils as well as with the poverty rates for the councils. Finally, Table 6

takes this investigation to the level of the 354 magisterial districts.[9] As mentioned above, the IES was not designed to be representative at this degree of disaggregation; this is reflected in the increased average goodness of fit. However, the increased sample size of the magisterial district regressions also allows for greater precision of the estimates as well as more confidence that the income and urban effects are not driven by a single observation. As before, the regressions show that difference between IES and census data are not invariant to the place where the sample was collected.

Table 4: Regression of goodness of fit on area of residence and mean expenditure (province/EA-type level)

Dependent variable: goodness of fit	**Fit between census income and IES expenditure**				**Fit between imputed census exp. and IES expenditure**			
	Mean expenditures		**Headcount indices**		**Mean expenditures**		**Headcount indices**	
	Coeff. (1)	**Coeff. (2)**	**Coeff. (3)**	**Coeff. (4)**	**Coeff. (5)**	**Coeff. (6)**	**Coeff. (7)**	**Coeff. (8)**
IES expenditure (,000)	0,061 (0,017)**	0,068 (0,024)**	0,083 (0,070)	0,009 (0,108)	0,004 (0,019)	0,033 (0,024)	-0,085 (0,039)*	-0,049 (0,050)
% former homelands	0,186 (0,060)**		0,831 (0,246)**		-0,015 (0,066)		-0,101 (0,134)	
% urban formal		-0,131 (0,068)*		-0,208 (0,303)		-0,096 (0,066)		-0,075 (0,141)
F(3,27)	6,50	3,94	7,02	2,45	0,35	1,05	6,97	6,80
Adjusted R^2	0,355	0,227	0,376	0,126	-0,070	0,005	0,374	0,367
N	31	31	31	31	31	31	31	31
Mean goodness of fit		0,187		0,905		0,103		0,185

Standard errors in parentheses.
*** denotes significance at the 5% level and**
**** at the 1% level.**

[9] We also explored specifications which included either the number of households in the district or the square root of this number to see if smaller MDs or Dcs had measurably greater deviation between the census and the IES data. The coefficients of cluster size were generally significant at the 10% level or less and with a sign consistent with the expectation that precision increased with the size of the cluster. However, neither the regression r-square values nor the magnitude of the coefficient of other variables were affected by the inclusion of the cluster size. Thus the regression reported in the tables do not include the number of households.

Table 5: Regression of goodness of fit on area of residence and mean expenditure (district council level)

Dependent variable: goodness of fit	**Fit between census income and IES expenditure**				**Fit between imputed census exp. and IES expenditure**			
	Mean expenditures		**Headcount indices**		**Mean expenditures**		**Headcount indices**	
	Coeff. (1)	**Coeff. (2)**	**Coeff. (3)**	**Coeff. (4)**	**Coeff. (5)**	**Coeff. (6)**	**Coeff. (7)**	**Coeff. (8)**
IES expenditure (,000)	0,102 (0,020)**	0,135 (0,024)**	0,169 (0,057)**	0,232 (0,079)**	0,070 (0,016)**	0,081 (0,019)**	0,030 (0,032)	0,092 (0,036)*
% former homelands	0,304 (0,076)**		1,36 (0,215)**		0,046 (0,060)		0,103 (0,121)	
% urban formal		-0,487 (0,106)**		-1,65 (0,357)**		-0,108 0,086)		-0,471 (0,162)**
F(3,41)	11,69	13,89	14,69	8,21	9,21	9,76	1,09	3,83
Adjusted R^2	0,422	0,468	0,483	0,330	0,359	0,374	0,006	0,162
N	45	45	45	45	45	45	45	45
Mean goodness of fit		0,243		0,888		0,176		0,177

Standard errors in parentheses.
*** denotes significance at the 5% level and**
**** at the 1% level.**

Table 6: Regression of goodness of fit on area of residence and mean expenditure (magisterial district level)

Dependent variable: goodness of fit	**Fit between census income and IES expenditure**				**Fit between imputed census exp. and IES expenditure**			
	Mean expenditures		**Headcount indices**		**Mean expenditures**		**Headcount indices**	
	Coeff. (1)	**Coeff. (2)**	**Coeff. (3)**	**Coeff. (4)**	**Coeff. (5)**	**Coeff. (6)**	**Coeff. (7)**	**Coeff. (8)**
IES expenditure (,000)	0,159 (0,010)**	0,171 (0,010)**	0,154 (0,023)**	0,146 (0,027)**	0,116 (0,010)**	0,128 (0,011)**	-0,016 (0,015)	0,002 (0,016)
% former homelands	0,282 (0,036)**		1,04 (0,084)**		0,167 (0,010)**		0,197 (0,056)**	
% urban formal		-0,360 (0,046)**		-0,910 (0,121)**		-0,257 (0,049)**		-0,337 (0,071)**
F(3,346)	93,5	92,4	57,3	23,8	43,0	46,74	6,79	10,1
Adjusted R^2	0,443	0,440	0,326	0,164	0,265	0,282	0,047	0,073
N	354	354	354	354	354	354	354	354
Mean goodness of fit		0,290		0,948		0,244		0,376

Standard errors in parentheses.
*** denotes significance at the 5% level and**
**** at the 1% level.**

To summarise: the income data collected in the census significantly understates the income or expenditure levels of the households measured by a detailed module in a household survey in South Africa. Similarly, the census data imply much higher rates of poverty than the IES data. Furthermore, this gap depends on the area of residence of the households. For households which live in areas classified as rural portions of former homelands or other rural areas, this gap is larger than that of those who live in urban areas. These two findings suggest that one should be very cautious in using the census income for policy purposes, as one is likely to over-estimate poverty in some areas, and possibly under-estimate it in others, with the bias being systematic. In the section that follows we propose an alternative measure also derived from the census with the help of the household survey.

Imputing expenditures in Census '96

As described in above, the methodology of imputing expenditures for each household in the census is conceptually simple, yet computationally intensive. It involves creating an association model between per capita household expenditure (or income) and household characteristics that are common to both the census and the household survey. After carefully constructing the variables in the exact same manner in each data set, we run a simple OLS regression of logarithmic per capita household expenditure on the other constructed variables that consist of household composition, education, primary occupation, quality of housing, and access to services. To avoid forcing the parameter estimates to be the same for all areas in South Africa, we run the regression separately for each of the nine provinces. The explanatory power of the nine regressions ranged from an R^2 of 0,6 (Northern Province) to 0,79 (Free State). As these are regressions based on household level observations, these values can be considered quite good. In Table 7, we show the results of our regression on the entire sample, i.e. covering all nine provinces in South Africa.

These regressions can be considered as components of an association model rather than a causal model. That is, the parameter estimates should not be interpreted as the effect of the explanatory variables on household expenditure. The parameters form a set of weights by which the household variables in census data are to be summed in order to get a measure of imputed expenditure. In effect, we use the set of parameter estimates to predict logarithmic per capita household expenditure for each household in the census in a manner quite similar to the construction of a basic needs indicator (BNI). However, while almost all BNIs that one can find in the literature use an *ad hoc* set of weights, our weights are informed by an association model from the household survey. Hentschel *et al.* (2000) shows that such *ad hoc* BNIs can lead to significant errors in spatial rankings compared to estimates of welfare, measured by household consumption.

Given the vector for the parameter estimates ß, and the vector of explanatory variables in the census X_c, the predicted log per capita expenditure for each household in the census is X_cß. This provides measures of per capita and total monthly expenditure for each household in the census. These can then be used to compare mean predicted expenditures from the census with point estimates for mean expenditures from the IES at the province (and geographical units of higher disaggregation) level.

Estimating standard errors is a bit more complicated. While the standard errors from the IES are the familiar estimates of the standard deviation based on sample theory, the issues of sample error does not exist in a census. However, there is a distribution around each imputation of expenditure for the census households. We will defer discussion of this until after the comparison between the point estimates of expenditures in the census and the IES estimates.

Table 7: Regression results by province

Variable	Western Cape	Eastern Cape	Northern Cape	Free State	KwaZulu-Natal
# of males aged 0-10	-0,153**	-0,125**	-0,121**	-0,221**	-0,079**
	(0,015)	(0,011)	(0,024)	(0,017)	(0,012)
# of males aged 11-20	-0,189**	-0,184**	-0,180**	-0,240**	-0,109**
	(0,017)	(0,012)	(0,028)	(0,018)	(0,013)
# of males aged 21-40	-0,111**	-0,158**	-0.148**	-0,175**	-0,070**
	(0,018)	(0,013)	(0,029)	(0,021)	(0,014)
# of males aged 41-65	-0,009	-0,073**	-0,095**	-0,097**	-0,058**
	(0,023)	(0,017)	(0,035)	(0,025)	(0,019)
# of females aged 0-10	-0,141**	-0,134**	-0,166**	-0,200**	-0,067**
	(0,016)	(0,011)	(0,025)	(0,018)	(0,012)
# of females aged 11-20	-0,179**	-0,163**	-0,214**	-0,251**	-0,105**
	(0,017)	(0,012)	(0,028)	(0,018)	(0,013)
# of females aged 21-40	-0,138**	-0,139**	-0,202**	-0,213**	-0,112**
	(0,019)	(0,014)	(0,032)	(0,020)	(0,014)
# of females aged 41-65	-0,185**	-0,161**	-0,183**	-0,252**	-0,154**
	(0,022)	(0,017)	(0,038)	(0,024)	(0,018)
# of individuals	-0,025**	-0,003	-0,030**	0,007	-0,039**
categorized as African	(0,007)	(0,005)	(0,008)	(0,008)	(0,006)
# of individuals	0,175**	0,128**	0,200**	0,214**	0,139**
categorized as white	(0,008)	(0,011)	(0,015)	(0,013)	(0,009)
Hh lives in a formal	-0,263**	0,158**	-0,124**	0,009	0,154**
dwelling	(0,040)	(0,021)	(0,053)	(0,027)	(0,025)
# of rooms per person	0,266**	0,245**	0,225**	0,197**	0,237**
	(0,010)	(0,008)	(0,016)	(0,010)	(0,010)
Hh owns the dwelling	0,183**	0,131**	0,128**	0,178**	0,181**
	(0,023)	(0,018)	(0,037)	(0,026)	(0,018)
Sanitary services available	0,207**	0,198**	0,285**	0,414**	0,289**
	(0,037)	(0,026)	(0,043)	(0,028)	(0,031)
Electricity for lighting	0,315**	0,261**	0,164**	0,266**	0,289**
available	(0,041)	(0,025)	(0,047)	(0,027)	(0,026)
Refuse removal 1 x week	0,024	-0,055**	0,148**	0,121**	-0,077**
	(0,031)	(0,023)	(0,046)	(0,031)	(0,028)
Telephone available	0,422**	0,334**	0,405**	0,244**	0,301**
	(0,027)	(0,029)	(0,045)	(0,032)	(0,026)
# of ind. who completed	0,054**	0,087**	0,081**	0,045**	0,048**
primary education	(0,011)	(0,007)	(0,017)	(0,012)	(0,008)
# of professionals	0,273**	0,511**	0,307**	0,433**	0,299**
	(0,016)	(0,016)	(0,034)	(0,019)	(0,014)
# of skilled labourers	0,141**	0,246**	0,198**	0,338**	0,169**
	(0,018)	(0,023)	(0,039)	(0,028)	(0,017)
Adjusted R^2	0,743	0,737	0,743	0,793	0,730
N	3213	5200	1419	3105	4933

Standard errors in parentheses.
*** denotes significance at the 5% level and**
**** at the 1% level.**
 # means number

Table 7: Regression results by province (continued)

Variable	North West	Gauteng	Mpumalanga	Northern Province
# of males aged 0-10	-0,124**	-0,099**	-0,055**	0,017
	(0,021)	(0,018)	(0,019)	(0,026)
# of males aged 11-20	-0,152**	-0,166**	-0,073**	-0,052*
	(0,021)	(0,019)	(0,020)	(0,027)
# of males aged 21-40	-0,099**	-0,053**	-0,035	-0,045
	(0,025)	(0,020)	(0,021)	(0,029)
# of males aged 41-65	-0,056*	-0,021	0,011	0,135**
	(0,031)	(0,025)	(0,028)	(0,035)
# of females aged 0-10	-0,123**	-0,110**	-0,032*	0,009
	(0,021)	(0,018)	(0,019)	(0,025)
# of females aged 11-20	-0,147**	-0,184**	-0,077**	-0,051*
	(0,022)	(0,020)	(0,020)	(0,026)
# of females aged 21-40	-0,162**	-0,160**	-0,095**	-0,083**
	(0,025)	(0,022)	(0,022)	(0,029)
# of females aged 41-65	-0,234**	-0,219**	-0,137**	-0,129**
	(0,030)	(0,025)	(0,028)	(0,034)
# of individuals categorized as African	-0,008	-0,080**	-0,077**	-0,130**
	(0,011)	(0,007)	(0,012)	(0,020)
# of individuals categorized as white	0,143**	0,104**	0,121**	0,033
	(0,016)	(0,008)	(0,016)	(0,026)
Hh lives in a formal dwelling	-0,199**	0,009	0,183**	0,230**
	(0,038)	(0,037)	(0,033)	(0,033)
# of rooms per person	0,264**	0,222**	0,234**	0,262**
	(0,014)	(0,011)	(0,014)	(0,017)
Hh owns the dwelling	0,233**	0,250**	0,274**	0,138**
	(0,029)	(0,024)	(0,027)	(0,039)
Sanitary services available	0,524**	0,282**	0,030	0,223**
	(0,040)	(0,054)	(0,040)	(0,047)
Electricity for lighting available	0,309**	0,308**	0,388**	0,255**
	(0,038)	(0,047)	(0,032)	(0,036)
Refuse removal 1 x week	-0,089**	0,126**	0,046	-0,189**
	(0,040)	(0,031)	(0,039)	(0,047)
Telephone available	0,319**	0,338**	0,152**	0,385**
	(0,042)	(0,026)	(0,040)	(0,050)
# of ind. who completed primary education	0,090**	0,070**	0,034**	0,117**
	(0,013)	(0,013)	(0,012)	(0,014)
# of professionals	0,425**	0,245**	0,356**	0,437**
	(0,024)	(0,015)	(0,024)	(0,025)
# of skilled labouers	0,214**	0,119**	0,209**	0,306**
	(0,031)	(0,021)	(0,026)	(0,037)
Adjusted R^2	0,716	0,699	0,709	0,600
N	2441	3247	2370	2634

Standard errors in parentheses.
*** denotes significance at the 5% level and**
**** at the 1% level.**
 # means number

How well do the imputed expenditure measures improve the fit between data sets? As already mentioned, the regression parameters reported in Table 7, allow us to derive a measure of expected household expenditure conditional on the quality of housing, services received and the composition of each household in the census. The average household expenditure from this imputation is R2 789 per month. This is only 6,4% below that in the IES. Thus, the difference between the imputed expenditures using census data and the IES expenditures is only a third as large as the difference between the average census income and the IES expenditures. While the average predicted value from an OLS regression will be the same as the average of the sample *from which it was derived*, this is not necessarily the case when fitting parameters to another data set. The fact that the predicted value corresponds to the average from the IES reflects the fact that the distribution of explanatory variables is similar in the two data sets. Furthermore, using the poverty line of R800 per household per month, we find an overall expected poverty incidence of 28,5% for South Africa, a figure which is virtually identical to the corresponding headcount index value (28,4%) from the IES.

The correlation coefficient between the provincial averages of census imputed expenditures and that from the IES expenditure is 0,97, and the Spearman rank correlation coefficient is 0,93 (Table 2). Similarly, the corresponding figures for the poverty measures (% of households with less than R800 per month) calculated from the two alternative data sources are 0,90 and 0,97, respectively. These are significant improvements over the previous figures that used census income. There is less improvement in the simple correlation coefficients for average income at lower levels of aggregation and, indeed, the correlation declines slightly at the MD level. However, the rank correlation for the averages do improve at all levels of aggregation. Even more germane to the objectives of this study, at all levels of aggregation, the expected poverty rates and poverty ranking correlate more closely with the corresponding observations in the IES than do the poverty rates using census income.[10]

Moreover, unlike the average income and poverty estimates based on the census data there is no systematic pattern in the difference between the imputed expenditures and the IES data. This is demonstrated by the last four columns of Tables 3-6. For example, in the last four columns in Table 3 there is no longer a significant effect of the areas of residence on the goodness of fit between the two measures. However, the coefficient for mean expenditure levels in each province remain significant and positive in the regressions for mean expenditures but not for poverty rates. Furthermore, the F statistics in both regressions are significant only at the 10% level and the explanatory power of each has dropped significantly. This is exactly what one would expect if there is only a weak relationship between area of residence and how closely the mean imputed census expenditure corresponds with expenditure from the household survey.

Table 4 indicates that when the unit of observation is averaged over the type of enumeration area in each province, the sign of the average expenditure is no longer consistently positive, and, as with Table 3, the type of residence no longer influences the goodness of fit. Note that the coefficient on dummy variable for the per cent of households residing in urban formal areas remains negative in the regression at the district and MD levels (Tables 5 and 6). However, the magnitude of this coefficient is greatly reduced compared to the regression results in columns 2 and 4, as are themeanvalues for the goodness of fit. As indicated above, a reduction in the goodness of fit measure indicates an improvement in the overall fit. Also as discussed, it should be borne in mind that the IES is not representative at this level and some of the observed imprecision may reflect sample error in that survey.

[10] If we look at the correlation of average income from the IES and average expenditures from that survey, we find that at the province and DC level the correlations are both 0,99. At the MD level the correlation is 0,96. For all three levels the rank correlations areabove 0,93.

Povertymapping using imputed expenditures from Census '96

Having established a closer correspondence of imputed expenditure in the census data to household expenditure in the IES than that of income from the census, we proceed to the primary objective for this paper, the construction of a poverty map for South Africa, using the imputed expenditures, at all levels of disaggregation. What we have done so far is this.[11] We have estimated 1st stage regressions for each province in the household survey:

$$lny_i = X_{i'}\beta + \varepsilon_i \qquad \varepsilon_i \_ N(0,\sigma^2) \tag{1}$$

where lny_i is the logarithm of per-capita consumption expenditure for household *i*, with independent variables X_i common to the IES and the census, and ε_i a random disturbance term. Using the predicted values of ß and σ, we can calculate our estimator of expected poverty for household *i* in the census by:

$$P_i^* = \hat{E}[P_i \mid X_i, \hat{\beta}, \hat{\sigma}] = \Phi\left(\frac{lnz - X_{i'}\hat{\beta}}{\hat{\sigma}}\right) \tag{2}$$

where P_i is the poverty for household *i*, z is the poverty line, and Φ indicates the cumulative standard normal distribution. Given that we aim to calculate the expected headcount poverty indicator, the value in (2) is simply the estimate of the probability that a household with observable characteristics X_i is poor. The intuition here is quite clear. Since the 1st stage regressions have an idiosyncratic component, there is always a non-zero probability that a household is poor however high its predicted expenditure may be. A weighted (by household size and sampling weights whenever available) average of these probabilities over any geographical unit would give us the expected percentage of poor individuals in that area. Thus, the predicted incidence of poverty P*, given the estimated model of consumption is

$$P^* = \hat{E}[P \mid X, \hat{\beta}, \hat{\sigma}] = \frac{1}{N}\sum_{i=1}^{N} n_i * \Phi\left(\frac{lnz - X_{i'}\hat{\beta}}{\hat{\sigma}}\right). \tag{3}$$

where N is the number of households in the area and n_i is the number of individuals in household *i*. These expected poverty rates are illustrated in Figure 1 and reported in the appendix. In Appendix Table 1, provinces are ranked by the expected headcount poverty rate in descending order, i.e. from poorest to the richest province. Appendix Tables 2 and 3 are sorted by province and then within the province, districts are ranked by the headcount index to illustrate the wide variation of expected poverty within each province.

For many uses of the imputed poverty rates or average imputed expenditures, such as making pair-wise comparisons, we need to calculate the error in our prediction in the census. To summarise the difference between our estimates of the expected poverty rates and the actual value of the poverty rates in population, we introduce the following notation. The interested reader should refer to Elbers *et al.* (2000), for a detailed discussion of the standard error calculations.

[11] The methodology employed here of calculating headcount indices from the imputed expenditures in the census is based on Hentschel *et al.* (1999). More details can be found in that paper.

Suppose that we denote the poverty in the population by $P(y) = P(X, ß, \varepsilon)$[12]. Since we do not know the actual vector of disturbances, ε^0, we estimate the expected value of this indicator, $E[P \mid X,\omega]$, where ω represents the vector of parameters $\{ß, \sigma^2\}$. Furthermore, when we construct an estimator for this expected value, we replace the unknown vector ω with consistent estimators, ϖ, from the 1st stage regressions described in equation (1) above. This yields $E[P \mid X,\varpi]$. Finally, since, for most of the FGT-class poverty measures and for all of the inequality measures, this expectation is analytically intractable, we use a method of computation that employs the actual distribution of the predicted log expenditures and a simulated distribution of the vector of disturbances, ε. We will denote this estimator by $E_s[P \mid X,\omega]$.

Hence, the difference between the value of the indicator, $P^0(y)$ and our estimator $E_s[P \mid X,\varpi]$ can be written as the following:

$$\begin{aligned} P^0(y) - E_s[P \mid X,\varpi] \;=\; & P^0(y) - E[P \mid X,\omega] + \\ & E[P \mid X,\omega] - E[P \mid X,\varpi] + \\ & E[P \mid X,\varpi] - E_s[P \mid X,\varpi] \end{aligned} \qquad \textbf{(4)}$$

This means that the error in our prediction can be broken down into three separate components. Elbers *et al*. call these three components the *idiosyncratic error*, the *model error*, and the *computation error*, respectively. The properties of each of these error components are discussed in their paper in detail. The standard errors of our expected poverty rates are small. In fact, for the levels of aggregation considered in our paper, the standard errors are such that most comparisons of expected poverty rates between provinces, district councils or magisterial districts yield differences that are statistically significant. These errors are reported in the appendix along with the expected headcount index figures for each of these administrative units. In the next section, we discuss possible extensions to our paper, and the likely implications of these extensions for our results.

The way forward: Stats SA and the World Bank

There are a number of important assumptions embedded in the methodology of Stats SA and the World Bank. The sensitivity of our results to these assumptions is an important issue that should not be overlooked. We discuss three main assumptions below. We also describe future work on sensitivity analysis.

First, we assume that the residuals from the 1st stage regressions are normally distributed. This is an assumption that is easy to test and easy to relax. Our preliminary analysis shows that our residuals do look normal when overlaid on a normal kernel density function, and in the cases where we do not pass the standard tests of normality, we find that this is due to the existence of a few outliers. [The tests for normality that we utilised are all readily implemented in STATA, such as sktest (skewness and kurtosis test), sfrancia (Shapiro-Francia test), and jb (Jarque-Bera test)]. After dropping a few of these observations (usually less than 1% of the total number of observations in a province) we cannot reject the null hypothesis that the residuals are normally distributed in each region. Furthermore, our results

[12] Poverty in the population depends on household size, but, without loss of generality, we have left it out of the discussion for simplicity of notation.

are not sensitive to the elimination of these few outliers from the sample in each region. Finally, one can easily relax the normality assumption by drawing from the pool of the residuals from the 1st stage regressions with replacement, rather than from a normal distribution. That is, one does not need to impose a certain distributional form on the residuals.

We also assumed initially that our residuals are homoskedastic. Further tests of this assumption showed us that in most of our nine regressions, the residuals are in fact heteroskedastic. To deal with heteroskedasticity, if it is there, we need to estimate its form and then draw residuals in the imputation stage accordingly. This is a fairly straightforward extension, especially if the assumption of normality holds, in which case the residuals can still be drawn from a uniform distribution for our simulations and then transformed to have an appropriate variance.

Finally, we assume that the disturbance term in our equation (1) is not correlated across households within a cluster, town, or a magisterial district. Ignoring the fact that a component of the disturbance term is shared within groups, our methodology would still yield unbiased estimates of expected poverty for small areas conditional on their observable characteristics, although the standard errors around these estimates would be underestimated (see Elbers *et al*, 2000). That is, for each town (or place name ormagisterial district, etc.), we do not know the true value of poverty but our expectation of poverty, given whatwecanobserve, isunbiased.

Incorporating interaction terms, other data sources (e.g. geographic information systems databases), and means of our current explanatory variables at the cluster (or town, or magisterial district) level into our regression models are all various ways to ameliorate possible 'small area effects'. We find in several instances that our explanatory variables are sufficiently informative that the assumption of independence of the disturbance term across households cannot be rejected. Elbers *et al*. (2000) find no random effects at the cluster level in rural areas of Ecuador, although they get significant and sizeable effects in urban areas. In similar work in Nicaragua, we found no sign of fixed or random effects at the 'municipio' *(municipality)* level in any of the seven regions, urban or rural.

Hence, what we plan to do next is to perform proper diagnostics to see whether our assumption of 'no small area effects' is violated. If so, and preliminary evidence shows that it very well might be, we will explore expanding our set of explanatory variables as described above. If the problem still persists, we will incorporate the component of the disturbance term that is due to a common cluster effect into our simulations in the imputation stage. In that case, the standard errors around our expected poverty rates will be larger than those that are reported in this paper, butwithout doing the diagnostics it is not possible to know howmuchlarger.

In addition to these issues of estimation, our future work will explore estimating other dimensions of poverty. It is possible that our results are sensitive to the choice of our poverty line and/or to the choice of the poverty indicator. In this paper, we have only concentrated on the expected poverty rates. There is no reason why this should be the preferred choice of any policy-maker when using poverty maps as targeting tools. The poverty gap measure, for example, is widely used because of its interpretation as the amount of money necessary to bring all the poor up to the poverty line. Poverty severity, another indicator in the general class of Foster-Greer-Thorbecke Index of poverty measures [FGT (α=2)], is another possibility. It is not clear that all of the rankings of magisterial districts in South Africa are robust to the choice of poverty indicator. Furthermore, we have chosen our household poverty line to be R800 permonth, because it has immediate policy relevance as described in the introduction of our

paper. Whether our rankings are sensitive to the choice of the poverty line is also an empirical question. We will explore both of these issues of robustness in a separate forthcoming paper.

Concluding discussion

We have shown that the income from the census data provides only a weak proxy for the average income or poverty rates at either the provincial level or at lower levels of aggregation. We have also shown a simple method of imputing expenditures using information in the IES. The values for household consumption obtained using the regression coefficients from the IES and the characteristics available in the census are plausible and provide a fair fit with the IES data. The expected poverty rates for each magisterial district based on this methodology are provided in the appendix.

Since we have attempted to validate the estimates with data in the IES, it might seem logical to simply use this data, and bypass the imputation. However, as discussed, the IES was not designed to be representative at lower levels of aggregation while the census is, by design, exhaustive (and, hence, representative) for any jurisdiction. That is, there is no sample error, although there may be non-sample errors in the manner in which complex information was captured. The imputations reported here are based on readily-observable characteristics of a household such as its composition as well as the characteristics of its housing.

Our purpose is not merely to explore measures of poverty at the province level. In many cases these districts are themselves heterogeneous and there is often the need to know the rates of poverty for lower tiers of administration or for sub-regions within a province. While we cannot *formally* test whether the imputations which we provide are more accurate than the original information on income in the census data for lower tiers of administration, the evidence that has been presented is supportive of the claim that the imputed consumption provides an unbiased measure of poverty. Thus, we believe that the measure of consumption constructed for each household can be aggregated at any level of administration that requires information on poverty at the local level. Indeed, because the technique provides a measure of consumption for each household in rather geographically defined enumeration areas, expected poverty estimates can be provided for aggregations that differ from that which existed at the time the census was undertaken. This assists in updating information as the process of decentralisation of government services progresses. Moreover, with improvements provided with geographic information systems, such mapping can be a valuable tool in prioritising government resource allocation.

References

Alderman, Harold. (1999*). Multi-tier Targeting of Social Assistance: Role of Inter-governmental Transfers.* World Bank. Processed.

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, pp. 328-336.

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, pp. 537-559.

Baker, Judy and Margaret Grosh. (1994). *Measuring the Effects of Geographic Targeting on Poverty Reduction*. World Bank. Living Standards Measurement Study Working Paper No. 99.

Bramley, G. and G. Smart. (1996). 'Modeling Local Income Distributions in Britain', *Regional Studies*, 30, pp. 239-255.

Deaton, Angus. 1997. *The Analysis of Household Surveys. A Microeconometric Approach to Development Policy*. Baltimore: Johns Hopkins University Press.

Elbers, C., P. Lanjouw, and J. Lanjouw. (2000). *Welfare in Towns and Villages. Micro-measurement of Poverty and Inequality*. Tinbergen Institute Working Paper. Forthcoming.

Greene, William. (1990). *Econometric Analysis*. New York: Macmillan Publishers.

Grosh, M. and E. Glinskaya. (1997). *Proxy Means Testing and Social Assistance in Armenia*, draft, Development Economics Research Group, World Bank.

Hentschel, J., J. Lanjouw, P. Lanjouw and J. Poggi. (2000). 'Combining Survey Data with Census Data to Construct Spatially Disaggregated Poverty Maps: A Case Study of Ecuador', *World Bank Economic Review*, forthcoming.

Inei. (1996). *Metodologia Para Determinar el Ingreso y la Proporcion de Hogares Pobres*, Lima.

Lanjouw, P., B. Milanovic, and S. Paternostro. (1999). *Economies of Scale and Poverty: the Impact of Relative Price Shifts During Transition*. World Bank Policy Working Paper No. 2009.

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

Paulin, G.D. and D.L. Ferraro. (1994). 'Imputing Income in the Consumer Expenditure Survey', *Monthly Labor Review*, December, pp. 23-31.

Ravallion, Martin. 1992. *Poverty Comparisons. A guide to concepts and methods.* Living Standards Measurement Study Working Paper No. 88. Washington DC: The World Bank.