ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Minot, Nicholas; Baulch, Bob

Working Paper

Poverty mapping with aggregate census data: What is the loss in precision?

WIDER Research Paper, No. 2004/38

Provided in Cooperation with:
United Nations University (UNU), World Institute for Development Economics Research (WIDER)

Suggested Citation: Minot, Nicholas; Baulch, Bob (2004) : Poverty mapping with aggregate census data: What is the loss in precision?, WIDER Research Paper, No. 2004/38, ISBN 9291906255, The United Nations University World Institute for Development Economics Research (UNU-WIDER), Helsinki

This Version is available at:
https://hdl.handle.net/10419/63619

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

United Nations University
WIDER
World Institute for Development Economics Research

Research Paper No. 2004/38

Poverty Mapping with Aggregate Census Data

What is the Loss in Precision?

Nicholas Minot[1] and Bob Baulch[2]

June 2004

Abstract

Spatially disaggregated maps of the incidence of poverty can be constructed by combining household survey data and census data. In some countries (notably China and India), national statistics agencies are reluctant, for reasons of confidentiality, to release household-level census data, but they are generally more willing to release aggregated census data, such as village- or district-level means. This paper examines the loss in precision associated with using aggregated census data instead of household-level data to generate poverty estimates. We show analytically that using aggregated census data will result in poverty rates that are biased downward (upward) if the rate is below (above) 50 percent and that the bias approaches zero as the poverty rate approaches zero, 50 percent, and 100 percent. Using data from Vietnam, we find …/…

Keywords: poverty mapping, small area estimation, Vietnam, aggregation bias
JEL classification: I3, O1, C4

Copyright © UNU-WIDER 2004

[1]Markets, Trade, and Institutions Division, International Food Policy Research Institute, Washington DC.
[2]Institute of Development Studies, University of Sussex.

This study has been prepared within the UNU-WIDER project on Spatial Disparities in Human Development, directed by Ravi Kanbur and Tony Venables, with Guanghua Wan.

UNU-WIDER acknowledges the financial contributions to the research programme by the governments of Denmark (Royal Ministry of Foreign Affairs), Finland (Ministry for Foreign Affairs), Norway (Royal Ministry of Foreign Affairs), Sweden (Swedish International Development Cooperation Agency—Sida) and the United Kingdom (Department for International Development).

ISSN 1810-2611 ISBN 92-9190-625-5 (internet version)

that the mean absolute error in estimating district-level poverty rates is 2.5 percentage points if the census data are aggregated to the enumeration-area level means and 3-4 percentage points if the data are aggregated to commune or district level. Finally, we propose a method for reducing the error using variances calculated from the census. Applying this approach to the Vietnam data, we show that this method can cut the size of the aggregation errors by around 75 percent.

Acknowledgements

We thank Phan Xuan Cam and Nguyen Van Minh for their help understanding the Vietnam census data, Michael Epprecht for help with GIS software and techniques, and Chris Elbers, Peter Lanjouw and Berk Özler for helpful methodological discussions. We also benefited from useful comments from two anonymous referees and participants at conferences at the United Nations University, Tokyo, 28-29 March 2003 and the Centre for the Study of African Economies, Oxford University, 18-19 March 2002. The usual disclaimers apply.

The World Institute for Development Economics Research (WIDER) was established by the United Nations University (UNU) as its first research and training centre and started work in Helsinki, Finland in 1985. The Institute undertakes applied research and policy analysis on structural changes affecting the developing and transitional economies, provides a forum for the advocacy of policies leading to robust, equitable and environmentally sustainable growth, and promotes capacity strengthening and training in the field of economic and social policy-making. Work is carried out by staff researchers and visiting scholars in Helsinki and through networks of collaborating scholars and institutions around the world.

www.wider.unu.edu *publications@wider.unu.edu*

UNU World Institute for Development Economics Research (UNU-WIDER)

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

Camera-ready typescript prepared by Lorraine Telfer-Taivainen at UNU-WIDER

Printed at UNU-WIDER, Helsinki

The views expressed in this publication are those of the author(s). Publication does not imply endorsement by the Institute or the United Nations University, nor by the programme/project sponsors, of any of the views expressed.

1 Introduction

Policymakers and researchers are interested in the geographic distribution of poverty for several reasons. First, knowledge of these patterns facilitates the targeting of programs designed, at least in part, to reduce poverty. Many countries use some form of geographic targeting in government programs such as credit, food aid, input distribution, healthcare, and education. Second, this information is useful in monitoring progress in addressing poverty and regional disparities. Third, it may provide some insight regarding the geographic factors associated with poverty, such as access to markets, climate, or topography.

In a growing number of countries, high-resolution poverty maps are now being produced using a relatively new two-stage approach. In the first stage, household survey data are used to estimate econometrically the relationship between poverty (or household expenditure) and a series of household characteristics, including household size and composition, education, occupation, housing characteristics, access to utilities, and ownership by consumer goods such as radios and bicycles. In the second stage, this relationship is applied to census data on the same household characteristics to calculate an estimate of the poverty rate[1] for some small geographic unit. Other poverty measures and indicators of income inequality can also be calculated, as well as standard errors of the estimates.

In an early application of this approach, Minot (1998, 2000) combined a probit regression on data from the 1993 Vietnam Living Standards Survey (VLSS) and district-level means of the household characteristics from the agricultural census in 1994 to estimate the ranking of the incidence of poverty across 543 rural districts. Hentschel et al. (1998, 2000) use household survey data and household-level census data to estimate disaggregated poverty rates for Ecuador. They show that with household-level census data it is possible to generate unbiased estimates of the poverty rate as well as estimates of the standard error of the poverty rates. In the first stage of this approach, the logarithm of per capita expenditure is regressed on household characteristics from a household survey. In the second stage, data on the same household characteristics from the census is used to predict per capita expenditures and derive various poverty (and inequality) measures. This method, further developed by Elbers et al. (2003), has been used to construct poverty maps for Cambodia, Guatemala, Mozambique, Malawi, Nicaragua, Panama, Peru, South Africa, and Vietnam—see Henninger and Snel (2002).

Researchers, however, do not always have access to household-level census data. The national statistics agencies in many countries are reluctant to release household-level

[1] In this paper, we use 'poverty rate' denoted by P_0, to refer to the percentage of people living in households whose per capita expenditure falls below the poverty line.

census data to researchers and international organizations, in part because of the issue of the confidentiality of the data. For example, China and India have each conducted a census within the past four years, but only district/county-level results are available to outside researchers. This means that household-level census data are not available to produce disaggregated poverty maps for 55 percent of the people who are living in extreme poverty worldwide.[2] In addition, the computational burden of processing census data, which may contain tens or even hundreds of millions of records, can be a challenge for even the most powerful desktop computers. When access to household census data or the computational burden of processing such data are constraining factors, one alternative is to use census data that have been aggregated to a higher level (such as the commune, district or province).[3] In this case, the researcher uses a database consisting of (for example) the district-level means of all the household characteristics for the second stage of the approach described above. An important question, however, arises: how much precision is lost in generating poverty maps from aggregate census data? If the errors are small, then reliable poverty maps can be produced for a wider range of developing countries. If the errors are large, then the use of aggregated data is not advisable and researchers should focus on getting access to household-level data.

This study uses recent household survey and census data from Vietnam to assess the loss in accuracy associated with the use of aggregated census data to estimate poverty instead of the original household-level data. The results of this analysis suggest that errors from using aggregated census data in the second stage of poverty mapping are, in the case of Vietnam, about 2-3 percentage points on average, if the level of aggregation is low. Furthermore, the paper shows analytically and empirically that the error is close to zero when the incidence of poverty is close to zero, close to 50 percent, or close to 100 percent. Results from using aggregated census data must be interpreted with caution, however, because this approach tends exaggerate differences between poor and less poor regions. We also propose a method to adjust for the aggregation bias and show that it can cut the mean errors by 75 percent.

The paper is divided into four sections. Section 2 describes the data and methods used to compare alternative measures of the incidence of poverty using household survey data and census data from Vietnam. Section 3 presents four types of results. First, we present an updated district-level map of poverty in Vietnam based on the best available data and methods. Then, we derive analytical results regarding the factors that affect the sign and relative magnitude of errors from the use of aggregate data. Next, we generate poverty estimates using Vietnamese census data that have been aggregated to different levels and compare the results to those obtained from the household-level census data. Finally, we

[2] According to calculations based on the World Development Indicators and using the US$1-a-day poverty line, China and India account for 55 percent of the world's poor (World Bank 2003).

[3] This approach has been used in Vietnam and in Gaza and the West Bank (see Minot 1998, 2000; Astrup and Dessus 2001).

propose and test a method for reducing the size of the errors associated with using aggregated census data. Section 4 summarizes the results and draws some implications for future research in poverty mapping.

2 Data and methods

2.1 Data

In this study, we use the 1998 VLSS and the 1999 Population and Housing Census. The VLSS was carried out by the General Statistics Office (GSO) of Vietnam with funding from the Swedish International Development Agency and the United Nations Development Program (UNDP), and with technical assistance from the World Bank. The survey was based on a stratified random cluster sample of 6,000 households, comprising 4,270 rural households and 1,730 urban households. The VLSS sample was based on ten strata: the rural areas of the seven regions and three urban strata (Hanoi and Ho Chi Minh City, other cities, and towns). For this analysis, we merge 'other cities' and 'towns' because the census data do not distinguish between these two strata.

The 1999 census was carried out by the GSO and refers to the situation as of 1 April 1999. It was conducted with the financial and technical support of the United Nations Family Planning Association and UNDP. Unit record data from the full census are not available, but a 33 percent sample was obtained from the GSO. The 33 percent sample was selected by GSO using systematic sampling of every third households, yielding a sample of 5.55 million households.

The VLSS and the census have a number of household variables in common: household size and composition, education of the head and spouse, housing characteristics, source of water, type of sanitation facility, ownership of three consumer goods (radios, televisions, and bicycles), and location of residence.

2.2 Methods

We begin with a description of the poverty mapping method when household-level census data are available. As mentioned above, the first step in implementing this approach is to use household survey data to estimate per capita expenditure as a function of a variety of household characteristics.[4] This typically takes the following semi-log form:

$$\ln(y_i) = X_i\beta + e_i \qquad (1)$$

where y_i is the per capita expenditure of household i, X_i is a $1\times k$ vector of characteristics of household i from the survey, β is a $k\times 1$ vector of estimated coefficients, and e_i is the

[4] Note that some 'household' characteristics (e.g., education or occupation of the household head) are based on the characteristics of individual members of the household. Some studies (for example, Bigman et al., 2000) also use community level characteristics in estimating per capita expenditures.

residual term.[5] To implement the regression analysis, we use the svyreg command in Stata, which takes into account the clustering, stratification, and other features of the sampling design. This command generates Huber/White/sandwich estimates of the standard error of the regression coefficients. We estimate separate models for rural and urban areas.[6]

The second step is to apply the regression equation to census data on the same household characteristics. If we are using household-level census data, this generates estimates of per capita expenditure for each household in the census. Hentschel et al. (2000) show that the expected value of the probability that household i is poor (P_i) can be described as follows:

$$E\left(P_i \mid X_i^C, \beta, \sigma\right) = \Phi\left[\frac{\ln(z) - X_i^C \beta}{\sigma}\right] \tag{2}$$

where $\Phi(\ )$ is the cumulative standard normal function, X_i^C is a vector of the same household characteristics taken from the census, β is a vector of the coefficients estimated in the first stage, z is the poverty line, and σ is the standard error of the regression from the first stage. If region contains N households labelled i= 1...N, the expected value of the poverty rate for the region, P, is simply the weighted average of the probabilities that the individual households are poor, where the weights are the share of the population in each household (m_i/M):

$$E\left(P \mid X_i^C, \beta, \sigma\right) = \sum_i \frac{m_i}{M} P_i = \sum_i \frac{m_i}{M} \Phi\left[\frac{\ln(z) - X_i^C \beta}{\sigma}\right] \tag{3}$$

In some cases, however, the statistics bureau of the government is not willing to release household-level census data but is willing to release aggregated data, such as the mean values of household characteristics for each district or village. The two studies of this type have used probit or logit regression models to predict whether households are poor or not instead of the semi-log model (Minot 1998, 2000, and Astrup and Dessus 2001). The mean values of the household characteristics in the census data are then inserted into the estimated probit/logit equation to estimate poverty for each aggregation unit in the census data (for example, for each district). This is not an unbiased estimate of poverty because the probit equation is non-linear. Using aggregate data ignores the variation in the household characteristics within each aggregation unit. For this reason, Minot (2000) used

---

5 Elbers et al. (2003) discuss a number of econometric issues related to this step, including the problems of heteroskedasticity and spatial autocorrelation. In the presence of these problems, our estimated coefficients would not be efficient, but the Huber/White/sandwich estimates of the standard errors used in this study are consistent under heteroskedasticity and take into account the effect of clustering and stratification on sampling error.

6 In Minot and Baulch (2002), we compare the poverty estimates obtained from rural/urban regression models to those obtained from eight stratum-level models. The urban/rural models gave a somewhat better fit (in terms of the value of R^2) and had more statistically significant coefficients. In any case, the difference in poverty estimates between the two approaches was quite modest, with provincial poverty rates (P_0) differing by an average of just 2.2 percentage points.

the results to rank districts by the incidence of poverty rather than reporting the estimated poverty rates. Even if we adopted the semi-log functional form in the first stage, the non-linearity of the cumulative standard normal function in Equation (3) would make it impossible to get an unbiased poverty estimate using aggregated census data.

Table 1: Summary of alternative methods to be compared

		Level of aggregation of poverty estimates		
		District	Province	Region
Level of aggregation of the census data	Household	Semi-log model	Semi-log model	Semi-log model
		Probit model	Probit model	Probit model
	Enumeration area	Semi-log model	Semi-log model	Semi-log model
		Probit model	Probit model	Probit model
	Commune	Semi-log model	Semi-log model	Semi-log model
		Probit model	Probit model	Probit model
	District		Semi-log model	Semi-log model
			Probit model	Probit model
	Province		Semi-log model	Semi-log model
			Probit model	Probit model
	Region			Semi-log model
				Probit model

Note: The underlined item represents the standard of comparison.

Source: See text.

In Section 3.1, we present the semi-log and probit regression models to 'predict' expenditure and poverty, respectively, based on household characteristics. Then we use the semi-log model and household-level census data to generate district-level estimates of the incidence of poverty in Vietnam. In Section 3.2, we use a second-order Taylor series expansion to provide an analytical expression for the error associated with using aggregate census data instead of household-level census data. This provides some information on the factors that influence the sign and relative magnitude of the error. In Section 3.3, we use data from Vietnam to examine the sensitivity of the poverty estimates to the choice of functional form in the first stage of the procedure and to the use of aggregate census data in the second stage. With regard to the functional form, we compare the results obtained from using a probit model and the semi-log model. With regard to the level of aggregation of the census data, we compare the estimates of the incidence of poverty (denoted by P_0) from the original household-level census data (considered the most accurate estimate) with estimates obtained from census data aggregated to the level of (a) the enumeration area, (b) the commune, (c) the district, (d) the province, and (e) the region.[7] The poverty estimates are calculated at four levels (district, provincial, regional, and national), though, of course,

[7] At the time of the 1999 census, Vietnam had 61 provinces, 614 districts, 10,714 communes and 166,481 enumeration areas (EAs).

the poverty estimates cannot be more disaggregated that the census data on which they are based.

Table 1 provides a summary of the methods being compared in this paper. The upper rows represent the (presumably) more accurate measures of poverty that use more disaggregated census data. The underlining indicates the standard of comparison used for each type of poverty estimate. The lower rows represent cruder approaches to estimating the incidence of poverty. For example, the third pair in the first column refer to the estimation of district-level poverty rates using commune averages of the indicators.

3 Results

3.1 District-level estimates of poverty in Vietnam

As described above, the first step in the poverty mapping procedure is to use household expenditure data to estimate per capita expenditure (or poverty) as a function of household characteristics. Table 2 provides the semi-log models of per capita expenditure in rural and urban areas using the VLSS. Table 3 presents the rural and urban probit models to predict which households are poor based on the same household characteristics. The second step is to apply the regression model to census data on the same household characteristics.

Figure 1 shows the district-level poverty rates obtained from applying the semi-log model to the household-level census data. The map indicates that poverty rates are over 80 percent in the districts bordering China to the north and Laos to the northwest. These areas are mountainous and have low population densities, poor transport infrastructure, and a high proportion of ethnic minorities. Many of the districts in the North Central Coast and the Central Highlands also have poverty rates between 40 percent and 80 percent. The Mekong Delta (at the southern tip) and the Red River Delta (on the northeastern coast) have poverty rates of 20 percent to 60 percent. These areas are favored by intensive irrigation of rice, fruits, and vegetables, good transportation networks, and proximity to the largest cities, Ho Chi Minh City and Hanoi. The districts with the lowest poverty rates (below 20 percent) are near Hanoi and in the southeast region. The southeast region includes Ho Chi Minh City, the largest and most commercially-oriented city in Vietnam. The rural areas around Ho Chi Minh City have become an important center for commercial agriculture and agro-industry. These patterns conform closely to the results from earlier studies (see World Bank 1995; Poverty Working Group 1999; and Minot 2000).

As mentioned above, with household-level census data it is possible to calculate standard errors and construct confidence intervals around the poverty estimates. The confidence intervals for the district-level poverty estimates in Figure 1 range from ±1.3 to ±22 percentage points, with a mean value of ±5.8 percentage points—see Minot et al. (2004) for more details.

Table 2: Semi-log regression models of per capita expenditure

	Rural	Model	Urban	Model
N	4269		1730	
R^2	0.536		0.550	
Variable	Coefficient	t	Coefficient	t
Household size	-0.0772	-19.5***	-0.0785	-8.1***
Percent elderly	-0.0831	-2.4**	-0.1026	-1.6
Percent children	-0.3353	-9.4***	-0.2368	-3.6***
Percent female	-0.1177	-3.5***	0.0386	0.5
Ethnic minority	-0.0765	-1.9*	0.0142	0.2
Head finished primary	0.0585	3.4***	0.0616	1.7*
Head lower secondary	0.0883	4.5***	0.0338	1.3
Head upper secondary	0.0884	3.3***	0.1368	3.2***
Head adv. tech. training	0.1355	4.2***	0.1603	3.5***
Head post-sec. education	0.2552	4.9***	0.1843	3.7***
No spouse	0.0173	1.0	0.0344	0.8
Spouse finished primary	0.0049	0.3	0.0642	1.9*
Spouse lower sec.	0.0132	0.6	0.0987	2.6***
Spouse upper sec.	0.0107	0.3	0.1912	2.7***
Spouse adv. tech. train.	0.0921	2.3**	0.1285	3.2***
Spouse post-sec. educ.	0.1571	2.7***	0.1752	3.1***
Manager/leader	0.1414	3.5***	0.2312	3.0***
Professional/technician	0.1350	3.3***	0.0576	1.2
Clerk/service worker	0.1362	3.4***	0.0357	0.9
Agriculture/forest/fish	-0.0163	-0.6	-0.0093	-0.2
Skilled laborer	0.0701	1.9*	0.0071	0.2
Unskilled laborer	-0.0586	-1.7*	-0.1599	-2.9***
Permanent house	-0.9228	-4.3***	-0.5194	-3.4***
Semi-permanent house	-0.3120	-3.6***	-0.4001	-3.8***
Area of perm. house	0.2958	5.7***	0.2001	5.4***
Area of semi-per. house	0.1180	5.2***	0.1403	4.6***
Electricity	0.0765	2.7***	-0.0026	0.0
Tap water	0.0828	1.4	0.2289	5.3***
Other clean water source	0.1157	4.4***	0.0340	0.6
Flush toilet	0.2700	5.5***	0.1311	2.2**
Latrine	0.0556	2.6**	0.0049	0.1
Owns television	0.2124	15.1***	0.2167	5.5***
Owns radio	0.1009	7.0***	0.1599	6.2***
Red River Delta	0.0314	0.6	0.0693	0.7
North Central Coast	0.0485	0.8	0.0445	0.6
South Central Coast	0.1373	2.2**	0.1460	1.9*
Central Highlands	0.1708	2.1**	(1)	
Southeast	0.5424	9.4***	0.4151	5.5***

table continues…

Mekong River Delta	0.3011	5.1***	0.1895	2.1**
constant	7.5327	108.7***	7.7538	64.7***

Note: Dependent variable is the log of per capita expenditure. Regression analysis uses the 'svyreg' command in Stata, taking into account sample design effects. The standard errors are the Huber/White/sandwich estimators. (1) Variable omitted because there are no urban Central Highland households in the VLSS sample. * coefficient is significant at the 10% level, ** at the 5% level, and *** at the 1% level.

Source: Semi-log regression analysis of 1998 Vietnam Living Standards Survey.

Table 3: Probit regression models of poverty

	Rural	Model	Urban	Model
N	4269		1730	
Correct prediction	77.2%		80.2%	
Variable	Coefficient	t	Coefficient	t
Household size	0.2646	14.3***	0.2016	5.2***
Percent elderly	0.2960	2.0*	0.6555	1.6
Percent children	1.1654	8.3***	1.9540	4.8***
Percent female	0.2654	1.8*	0.6060	1.8*
Ethnic minority	0.3480	2.5**	-1.4063	-2.7***
Head finished primary	-0.1838	-2.7***	0.0492	0.3
Head lower secondary	-0.2715	-3.0***	-0.1834	-0.7
Head upper secondary	-0.2188	-1.9*	-0.9618	-2.1**
Head adv. tech. training	-0.1901	-1.3	-0.1838	-0.8
Head post-sec. education	-0.9608	-2.9***	(1)	
No spouse	0.0570	0.7	-0.0772	-0.4
Spouse finished primary	-0.0304	-0.4	-0.3579	-1.6
Spouse lower sec.	0.0401	0.4	-0.1744	-0.8
Spouse upper sec.	0.0946	0.6	-0.4886	-1.4
Spouse adv. tech. train.	-0.3949	-1.9*	-0.9367	-2.8***
Spouse post-sec. educ.	-1.2828	-3.3***	(1)	
Manager/leader	-0.7908	-3.2***	(1)	
Professional/technician	-0.5138	-2.6***	-0.6283	-1.2
Clerk/service worker	-0.4315	-2.4**	0.2873	1.2
Agriculture/forest/fish	-0.0490	-0.4	0.0570	0.3
Skilled laborer	-0.2490	-1.7*	0.0747	0.3
Unskilled laborer	0.0926	0.7	0.6134	3.2***
Permanent house	2.0174	2.3**	1.1957	0.8
Semi-permanent house	0.7057	1.8*	0.5558	0.9
Area of perm. house	-0.6689	-2.9***	-0.4698	-1.2
Area of semi-per. house	-0.2889	-2.7***	-0.2922	-1.7*
Electricity	-0.1990	-1.9*	-0.0948	-0.2
Tap water	-0.1337	-0.4	-0.4582	-2.2**
Other clean water source	-0.3644	-3.4***	0.2702	1.2
Flush toilet	-0.6064	-2.7***	-0.4153	-1.5
Latrine	-0.0802	-1.1	-0.1649	-1.0
Owns television	-0.6760	-11.9***	-0.7611	-3.6***
Owns radio	-0.2998	-5.1***	-0.1169	-0.8

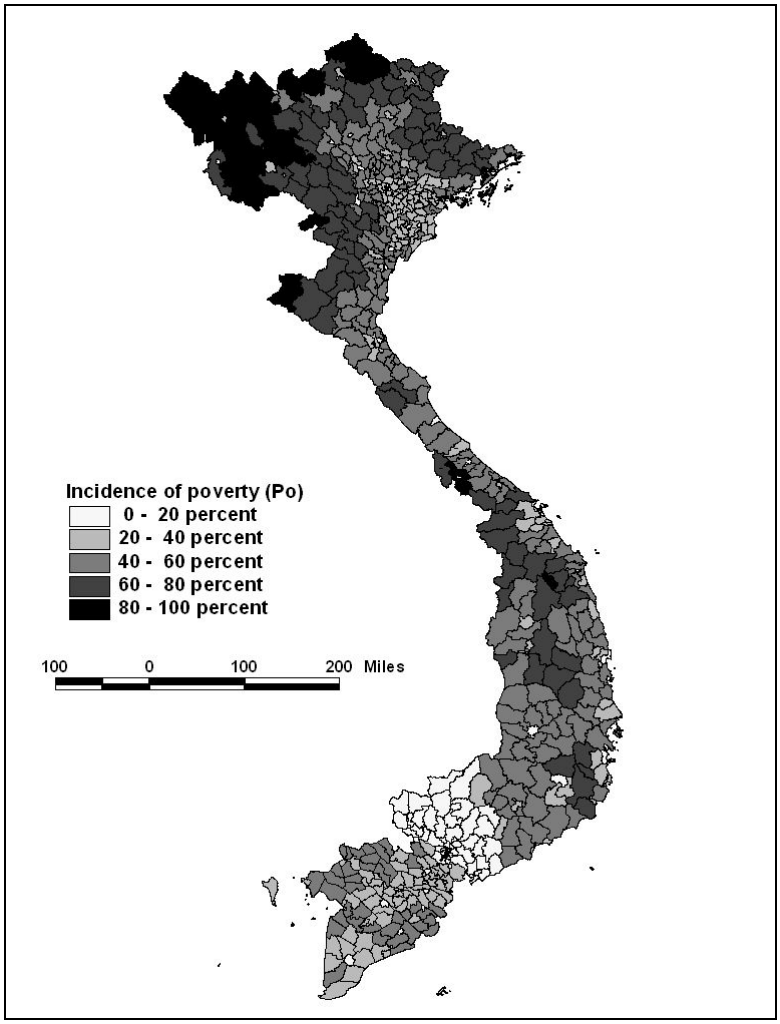
table continues…

Red River Delta	-0.1269	-0.7	0.5038	1.8*
North Central Coast	-0.1736	-0.8	0.5167	2.0**
South Central Coast	-0.5567	-2.8***	-0.0825	-0.3
Central Highlands	-0.8070	-2.9***	(2)	
Southeast	-1.6979	-7.9***	-0.6654	-1.7*
Mekong River Delta	-0.9502	-4.3***	-0.1820	-0.6
constant	-0.2816	-1.1	-1.8916	-3.6***

Note: Dependent variable is 1 if the household is poor and 0 if not. Regression analysis uses the 'svyprobt' command in Stata, taking into account sample design effects. The standard errors are the Huber/White/sandwich estimators. (1) Variable omitted because it perfectly predicts not being poor. (2) Variable omitted because there are no urban Central Highland households in VLSS sample. Coefficient is significant at the 10% level, ** at the 5% level, and *** at the 1% level.

Source: Probit regression analysis of 1998 Vietnam Living Standards Survey.

Figure 1: District-level estimates of poverty (P_0)



Source: Authors' configuration.

3.2 Determinants of the errors of aggregation

Suppose that we can only obtain district-level means of the household characteristics from the census and we wish to calculate district-level poverty rates. The sign and magnitude of

the error associated with using aggregate census data instead of household-level census data can be estimated using a second-order Taylor expansion as follows:[8]

$$\frac{1}{N}\sum_{i}\Phi\left[\frac{\ln(z)-X_i^C\beta}{\sigma}\right]\cong\Phi\left[\frac{\ln(z)-\overline{X}^C\beta}{\sigma}\right]+\frac{1}{2}\text{var}\left(\frac{\ln(z)-X_i^C\beta}{\sigma}\right)\Phi''\left[\frac{\ln(z)-\overline{X}^C\beta}{\sigma}\right] \quad (4)$$

where the index i refers to households in the district, N is the number of households in the district, and $\overline{X}^C$ is the vector of district-level means of the household characteristics. The left-hand side of this equation represents the incidence of poverty as estimated from household-level census data (X_i^C), as described in Section 2.2. The first term on the right-hand side is the (less accurate) estimate of the incidence of poverty rate obtained from the aggregated census data ($\overline{X}^C$). The second term on the right side is the approximate error associated with using aggregate census data rather than household-level census data.[9] This error is a function of the variance of estimated log per capita expenditure within the aggregation region and the second derivative (or 'curvature') of the cumulative standard normal function at the means of the aggregation region.[10]

This equation has three implications for the error associated with using aggregate census data in poverty mapping. First, the variance is always positive and since the second derivative of the cumulative standard normal function is positive (negative) when the value of the function is below (above) 0.5, so poverty estimates based on aggregated data will underestimate poverty in regions with poverty rates below 50 percent and overestimate poverty in regions with poverty rates above 50 percent. In other words, if a country has regions with poverty rates below 50 percent and others with rates above 50 percent, using aggregate data to produce a poverty map will exaggerate the differences in poverty between the two sets of regions. Second, since the curvature of the cumulative standard normal function is zero in the center of the center of the curve and approaches zero at the two tails of the function, the error term approaches zero when the incidence of poverty is close to 0 percent, 50 percent, and 100 percent. Third, the magnitude of the error is proportional to the variance of the estimated log per capita expenditure within the geographic unit of aggregation. In the extreme, if there were no variation across households, there would be no error associated with using aggregate data. If we assume, as is plausible, that the variance in household characteristics is greater in larger geographic units, then aggregation over small units (such as a district) would produce smaller errors than aggregation over larger units (such as a province).

---

[8] The derivation of Equation (4) can be found in Appendix A.

[9] This is the *approximate* error because we started with the Taylor series expanded only to the second order. A more precise estimate of the error would take into account the third and higher-order terms in the series.

[10] Note that the poverty line (z) and the standard error of the regression (σ) are generally constant across the relatively small geographic units for which the incidence of poverty is estimated.

Although these results provide us with some information about the factors that determine the direction and relative magnitude of the errors associated with using aggregated census data in poverty mapping, they do not give us a sense of the absolute size of the errors. For example, errors of less than one percentage point would be considered negligible for most purposes, while errors of more than ten percentage points would be considered unacceptable to most users. In the next section, we use data from Vietnam to measure the error associated with using aggregated census data to produce estimates of the incidence of poverty.

3.3 Empirical comparison of alternative methods

As shown in Table 1, we can estimate the incidence of poverty at different levels of aggregation using census data aggregated to different levels. For example, we can calculate the incidence of national and regional poverty using the original household-level census data on the household characteristics, and compare these results with those produced from household characteristics averaged at different levels: the enumeration area (EA), commune, district, province, and region. Furthermore, we can use either the probit model or the semi-log model in the first stage. This yields twelve sets of estimates for national and regional poverty, as shown in Table 4.

The national poverty rate, estimated using household-level census data and the semi-log model, is 36.7 percent. Using aggregate census data, the estimates are 2 to 2.5 percentage points lower, ranging from 34.1 to 34.7 percent. Looking at the regional poverty estimates, when aggregated census data is used, the poverty rate is overestimated in the poorest region (Northern Uplands) and underestimated in the least poor regions (the two urban strata, the two deltas, and Rural Southeast). These results are consistent with equation (4) which predicts that aggregate data will underestimate (overestimate) poverty when the rate is below (above) 50 percent. On the other hand, using the semi-log model combined with the EA, commune, district or provincial level means, the ranking of regions by poverty rate is very similar to that with the household-level data. All twelve methods agree that the rural Northern Uplands region is the poorest and that Hanoi/Ho Chi Minh City is the least poor.

Table 5 compares the results from the semi-log model with household census data (column 1 in Table 4) and those of other methods (columns 2 to 12 in Table 4). The use of aggregate data appears to bias downward the regional poverty rates by between 2 and 3 percentage points on average, for the reasons mentioned above. As expected, the mean absolute error rises with the degree of aggregation in the census data. For example, the mean absolute error associated with the semi-log model rises from 2.1 percentage points

Table 4: Regional poverty estimates using different methods

	Household-level data		EA-level means		Commune means		District means		Provincial means		Regional means	
	Semi-log	Probit	Semi-log	Probit	Semi-log	Probit	Semi-log	Probit	Semi-log	Probit	Semi-log	Probit
Hanoi and HCMC	0.047	0.048	0.018	0.014	0.012	0.009	0.010	0.007	0.009	0.006	0.009	0.005
Other urban areas	0.155	0.135	0.114	0.080	0.099	0.065	0.094	0.060	0.084	0.049	0.073	0.039
Rural N Uplands	0.606	0.633	0.614	0.644	0.619	0.651	0.627	0.661	0.637	0.673	0.664	0.710
Rural Red R Delta	0.380	0.387	0.355	0.359	0.350	0.355	0.347	0.353	0.346	0.351	0.345	0.350
Rural N C Coast	0.506	0.523	0.501	0.519	0.502	0.520	0.503	0.522	0.510	0.532	0.510	0.532
Rural S C Coast	0.479	0.452	0.468	0.437	0.467	0.435	0.468	0.436	0.472	0.438	0.471	0.438
Rural C. Highlands	0.536	0.486	0.538	0.482	0.541	0.479	0.546	0.480	0.550	0.482	0.552	0.482
Rural Southeast	0.126	0.132	0.081	0.082	0.068	0.069	0.063	0.063	0.058	0.059	0.054	0.055
Rural Mekong Delta	0.396	0.405	0.370	0.381	0.363	0.375	0.361	0.373	0.359	0.372	0.356	0.369
Vietnam	0.367	0.369	0.347	0.346	0.342	0.341	0.342	0.341	0.342	0.341	0.343	0.344

Source: Estimated from 1998 VLSS and 33% sample of 1999 Population and Housing Census.

Table 5: Errors in regional poverty estimates using different methods

	Household-level data	EA-level means		Commune means		District means		Provincial means		Regional means	
	Probit	Semi-log	Probit	Semi-log	Probit	Semi-log	Probit	Semi-log	Probit	Semi-log	Probit
Bias	-0.003	-0.019	-0.026	-0.023	-0.030	-0.023	-0.030	-0.023	-0.030	-0.022	-0.028
Mean absolute error	0.018	0.021	0.037	0.027	0.043	0.030	0.046	0.034	0.050	0.039	0.057
Median absolute error	0.017	0.025	0.038	0.030	0.043	0.032	0.043	0.034	0.041	0.039	0.042
Mean squared error	0.001	0.001	0.002	0.001	0.002	0.001	0.003	0.002	0.003	0.002	0.004
Distribution of errors											
0-5 percentage points	100%	100%	78%	78%	67%	78%	56%	78%	56%	67%	56%
5-10 percentage points	0%	0%	22%	22%	33%	22%	44%	22%	33%	33%	22%
Over 10 percentage points	0%	0%	0%	0%	0%	0%	0%	0%	11%	0%	22%
Correlation coefficient	0.993	0.999	0.992	0.999	0.990	0.998	0.989	0.997	0.987	0.995	0.982
Rank correlation coefficient	0.983	1.000	0.967	1.000	0.967	1.000	0.967	1.000	0.967	1.000	0.967

Note: Errors are calculated relative to the poverty rates obtained using semi-log regression and household-level census data. Statistics are calculated giving equal weights to each region, so the bias is not equal to the difference in national poverty rates.

Source: Estimated from 1998 VLSS and 33% sample of 1999 Population and Housing Census.

for the EA-level aggregation to 3.0 percentage points for district means, and 3.9 for regional means. The error associated with the probit models is around 1.6 percentage points higher than that associated with the semi-log models at the same level of aggregation. Rows 5 through 7 of Table 5 show the percentage distribution of the regions according to the size of the error. When poverty is estimated using EA-level means and the semi-log model, the errors for all nine regions are less than 5 percentage points. Even when regional poverty rates are inferred from *regional* means in the household characteristics, the errors are less than 5 percentage points for six of the nine regions. The last two rows of Table 5 reveal a high degree of correlation across poverty estimates and high correlations of the regional rankings they generate.

The ability of aggregated census data to estimate regional poverty rates is interesting but perhaps less relevant than their ability to estimate provincial and district-level poverty rates. The real advantage of combining survey and census data is to be able to map poverty at these more disaggregated levels. Table 6 presents a summary of the errors in estimating the incidence of provincial poverty. Once again, the aggregated data introduce a small downward bias in the headcount incidence of poverty. The bias remains relatively constant, between -1 and -2 percentage points, regardless of the degree of aggregation of the census data. On the other hand, the mean absolute error is 2.2 percentage points for the semi-log model with EA-level means rising gradually to 3.6 percentage points for the semi-log model with provincial means. The percentage of provinces with absolute errors of less than 5 percentage points falls from 100 percent with the semi-log model and EA-level means to 77 percent with the semi-log model and provincial means. The probit models have mean absolute errors about 1 percentage point greater than the semi-log models using the same level of aggregation.

The four diagrams within Figure 2 plot the provincial poverty estimates based on household-level census data (on the horizontal axis) against estimates based on different levels of aggregation for the census data (on the vertical axis), using the semi-log model for both. The errors appear as deviations from the diagonal line. Panel (a) shows the close correspondence between provincial poverty estimates derived from household-level census data and those derived from EA-level means of the census data. Panels (b), (c) and (d) illustrate the progressively larger errors as the level of aggregation moves up from commune-level to district-level to provincial means. The elongated S-shaped pattern confirms the pattern predicted from equation (4) and discussed above, in which aggregated data result in an underestimate of poverty for less poor regions and an overestimate of poverty for the poorest regions. The goodness-of-fit multiple correlation coefficient (R^2) exceeds 0.99 for all four pairs of variables. This implies that over 99 percent of the variation in the provincial poverty rates can be 'explained' by the means of the household characteristics in the census data. Similarly, if the poverty estimates are ranked and their ranks compared, the (Spearman) rank correlation coefficient exceeds 0.995 for all pairs.

Figure 2: Provincial poverty estimates from aggregated census data compared to estimates from household-level census data

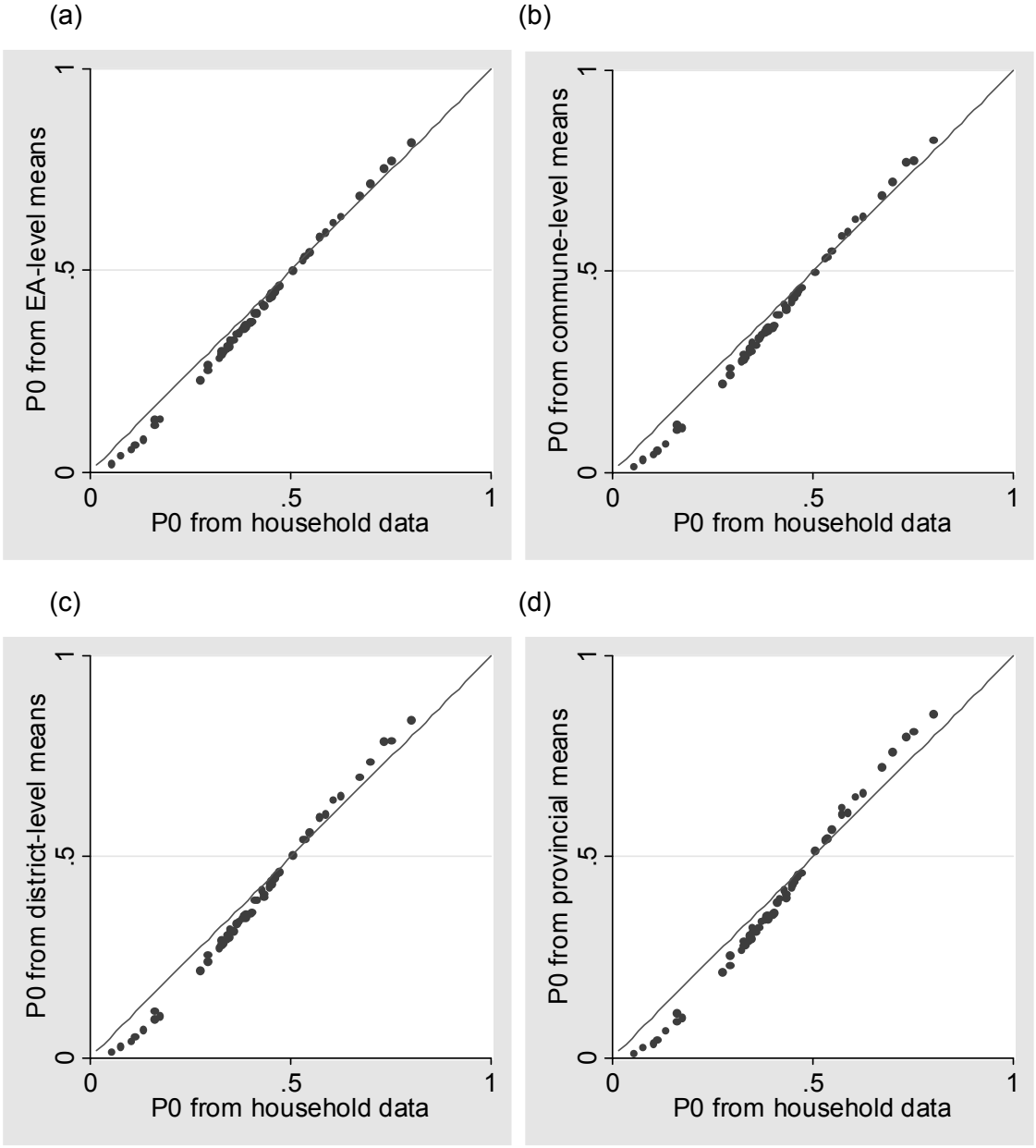


Table 7 and Figure 3 compare the district-level poverty estimates obtained from household-level census data and those obtained from aggregated census data. As would be expected, given the smaller sample size, the bias, mean and median errors are somewhat larger than the errors in provincial poverty estimates reported in Table 6. The bias never exceeds 2 percentage points, and the mean absolute error ranges from 2.0 to 4.8 percentage points, depending on the level of aggregation and the model. However, Figure 3 reveals the same pattern of errors as Figure 2, in which the incidence of poverty is exaggerated for the poorest districts and understated for the least poor districts. As explained above, this is due to the change in sign of the curvature of the cumulative standard normal function when the incidence of poverty rises above 50 percent. Again, the elongated S-shape is more pronounced when the level of aggregation is higher, in panel (c) which measures the

Table 6: Errors in provincial poverty estimates using different methods

	Household-level data	EA-level means		Commune means		District means		Provincial means	
	Probit	Semi-log	Probit	Semi-log	Probit	Semi-log	Probit	Semi-log	Probit
Bias	0.001	-0.016	-0.018	-0.020	-0.021	-0.019	-0.020	-0.018	-0.018
Mean absolute error	0.015	0.022	0.031	0.028	0.037	0.032	0.041	0.036	0.046
Median absolute error	0.011	0.021	0.028	0.027	0.037	0.030	0.042	0.035	0.045
Mean squared error	0.000	0.001	0.001	0.001	0.002	0.001	0.002	0.002	0.003
Distribution of errors									
0-5 percent	93%	100%	84%	90%	74%	87%	64%	77%	56%
5-10 percent	7%	0%	16%	10%	25%	13%	34%	23%	43%
over 10 percent	0%	0%	0%	0%	2%	0%	2%	0%	2%
Correlation coefficient	0.991	0.999	0.990	0.999	0.989	0.998	0.988	0.997	0.987
Rank correlation coefficient	0.981	0.999	0.983	0.999	0.982	0.998	0.981	0.999	0.982

Note: Errors are calculated relative to the poverty rates obtained using semi-log regression and household-level census data. Statistics are calculated giving equal weights to each province, so the bias is not equal to the difference in national poverty rates.

Source: Estimated from 1998 VLSS and 33% sample of 1999 Population and Housing Census.

Table 7: Errors in district poverty estimates using different methods

	Household-level data	EA-level means		Commune means		District means	
	Probit	Semi-log	Probit	Semi-log	Probit	Semi-log	Probit
Bias	0.000	-0.014	-0.017	-0.017	-0.020	-0.015	-0.018
Mean absolute error	0.020	0.025	0.036	0.032	0.043	0.038	0.048
Median absolute error	0.014	0.024	0.032	0.031	0.040	0.035	0.044
Mean squared error	0.001	0.001	0.002	0.001	0.003	0.002	0.003
Distribution of errors							
0-5 percent	91%	96%	73%	83%	62%	74%	56%
5-10 percent	9%	4%	25%	16%	35%	26%	38%
Over 10 percent	1%	0%	3%	0%	3%	0%	6%
Correlation coefficient	0.991	0.999	0.990	0.998	0.989	0.997	0.988
Rank correlation coefficient	0.987	1.000	0.988	0.999	0.987	0.999	0.987

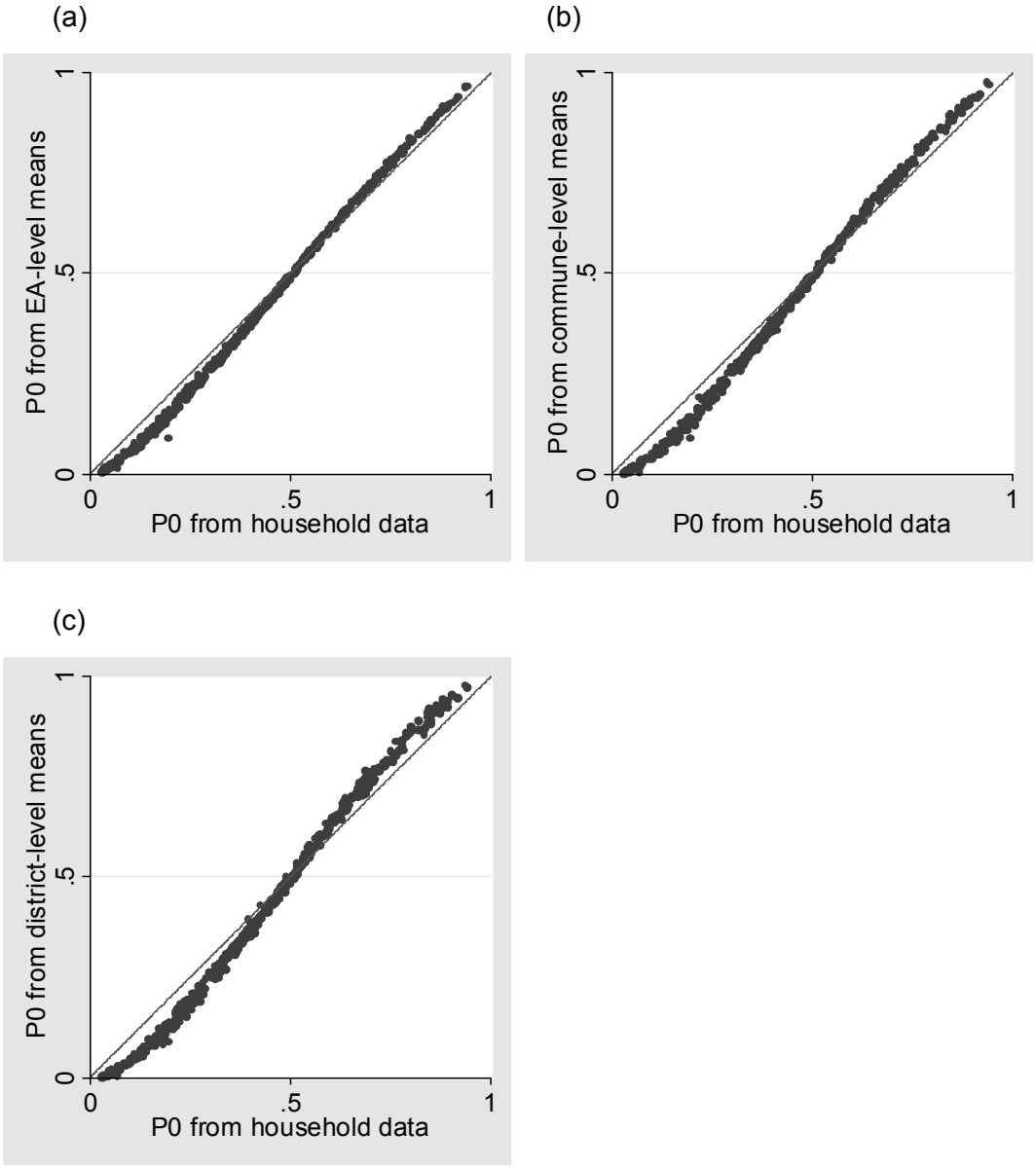
Note: Errors are calculated relative to the poverty rates obtained using semi-log regression and household-level census data. Statistics are calculated giving equal weights to each district, so the bias is not equal to the difference in national poverty rates.

Source: Estimated from 1998 VLSS and 33% sample of 1999 Population and Housing Census.

17

accuracy of district poverty estimates derived from district-level means of the census data on household characteristics.

Figure 3: District-level poverty estimates from aggregated census data compared to estimates from household-level census data



3.4 Improving poverty estimates derived from aggregate census data

In the previous section, we showed that empirical estimates of the errors associated with using aggregate census data were consistent with our expectations based on Equation (4). In this section, we show that Equation (4) can be used to improve poverty estimates derived from aggregate census data. As discussed above, the second term on the right-hand side of Equation (4) approximates the gap between the poverty estimate obtained from household-level census data and the poverty estimate obtained from aggregate census data. This term includes the variance of the normalized predicted log per capita expenditure,

var((ln(z)- $X_i^C\beta)/\sigma$), and the curvature of the cumulative standard normal curve $\Phi''((\ln(z)-\overline{X}^C\beta)/\sigma)$. The curvature component is simply the slope of the standard normal density curve and is easily calculated using numerical methods and the aggregated census data. The variance component, however, requires information at the household level.

One approach is to use the household survey data to calculate the variance component.[11] We can adjust the regional poverty estimates obtained from census data aggregated to the regional level by using regional variances calculated from the VLSS. For example, the estimate of poverty in the rural Northern Uplands using regional means is 0.664, almost 6 percentage points above the estimate of 0.606 based on household-level data. But the variance component is 0.622 and the curvature is -0.154, so the adjusted poverty rate is 0.664+(0.5)(0.622)(-0.154) = 0.616, which is just 1 percentage point above the poverty estimate based on the household survey data.

As shown in Table 8, the mean absolute error across the nine regions falls from 3.9 percentage points to 1.1 percentage points or 75 percent. Unfortunately, the variance cannot be reliably estimated at the provincial or district level because of the limited sample size of the VLSS, a problem likely to occur with any household survey. If we try to apply the regional variance to correct provincial and district-level poverty estimates within that region, the results are much less impressive: the adjustment reduces the mean absolute error by just 18-20 percent (see Table 8). Thus, the use of household survey data to estimate the variance component does not seem to be very promising.

An alternative approach is to ask the census authorities to calculate the variance component at some (preferably low) level of aggregation. Even if the census authorities are reluctant to release household-level data for reasons of confidentiality, they may still be willing to calculate this variance at the level of (for example) the enumeration area (EA). Under most circumstances, it is sufficient to obtain the variance of $X_i^C\beta$ in order to make these adjustments.[12]

The lower half of Table 8 shows that adjusting poverty estimates using EA-level variances from the census data can dramatically improve the precision of the poverty estimates derived from aggregate census data. In this case, EA-level poverty estimates are calculated using the EA-level means of the household characteristics and the EA-level variance of $X_i^C\beta$. The results are then aggregated to the district, provincial, and regional level and compared to the estimates obtained from household-level census data. These corrections reduce the mean absolute error for regional, provincial, and district-level poverty estimates

---

11 We are grateful to an anonymous reviewer for suggesting this approach.

12 Within a single regression domain (in which β and σ are constant) and within a single poverty-line domain (in which z is constant), it is only necessary to ask census authorities to calculate var($X_i^C\beta$) because $(1/\sigma^2)$var($X_i^C\beta$) = var[(ln(z)- $X_i^C\beta)/\sigma$].

from 2.1-2.5 percentage points to 0.5-0.6 percentage points. Thus, when household-level census data are not available, information about the variance of $X_i^c\beta$ is valuable in sharply reducing the errors associated with using aggregate census data.

Table 8: Effect of adjusting poverty estimates for errors of aggregation

Type of adjustment[1]	Level of poverty estimate	Level of aggregation of census data	Mean absolute error[2] of poverty estimate		Reduction in mean absolute error due to adjustment, %
			without adjustment	with adjustment	
Variance[3] at the regional level from household survey data	Region	Region	0.039	0.010	75
	Province	Province	0.036	0.029	18
	District	District	0.038	0.030	20
Variance[3] at the level of the enumeration area from census data	Region	Enumeration areas	0.021	0.005	77
	Province	Enumeration area	0.022	0.005	78
	District	Enumeration area	0.025	0.006	74

Notes: [1]'Adjustment' refers to the method of correcting for errors of aggregation using Equation 4. [2]'Mean absolute error' refers to the average absolute value of the difference between the poverty estimates obtained from household-level census data and those obtained from aggregated census data. [3]'Variance' refers to $var((\ln(z)-X_i^C\beta)/\sigma)$ from Equation 4.

Source: Estimated from 1998 VLSS and 33% sample of 1999 Population and Housing Census.

4 Summary and conclusions

This paper explores the errors associated with using aggregated census data instead of household-level census data in carrying out poverty mapping analysis. The issue arises because national statistics agencies in many developing countries (in particular, China and India) are reluctant to release household-level census data. Our analytical results suggest that the use of aggregated data will underestimate the incidence of poverty when the rate is below 50 percent and overestimate it where the rate is above 50 percent. The magnitude of the error varies with the estimated incidence of poverty, being smallest when the poverty rate is close to zero, 50 percent, and 100 percent. Furthermore, the error is proportional to the variance in estimated log per capita expenditure within the aggregated geographic units.

Empirical results using data from Vietnam indicate that, if census data are aggregated to the level of the enumeration area (each of which has about 85 households), the errors in estimating the incidence of poverty are relatively small, averaging between 2.1 and 2.5 percentage points for national, regional, and provincial estimates of poverty. Furthermore,

when the poverty rate is estimated using EA-level means of the census data, all 61 provinces and 96 percent of the 614 districts have errors of less than 5 percentage points. Not surprisingly, errors are larger when the level of aggregation is greater. Using census data aggregated to the level of communes or districts produces mean absolute errors of 2.8 to 3.8 percentage points. The study also compared the use of the semi-log regression model with that of the probit regression model. The incidence of poverty estimated from the probit model differed from that obtained from the semi-log model by about 1.0 percentage point for district-level and provincial poverty estimates.

Finally, we propose a method of adjusting the poverty estimates derived from aggregated census data. In particular, we show that information on the variance of $X_i^C\beta$ in the census data can be used to adjust the poverty estimates from aggregate data. This method cuts the mean absolute error associated with using aggregate census data by approximately three-quarters.

What are the implications of these results for other studies that combine household survey data and census data to produce high-resolution poverty maps? Clearly, the best option is to carry out the analysis with household-level census data. Not only does this generate more accurate estimates of the incidence of poverty, but it allows the estimation of various other measures of poverty and inequality (as well as estimates of standard errors of these measures) all of which are difficult to estimate with aggregated census or grouped household survey data (see Chen et al. 1991; Elbers et al. 2003).

At the same time, the results presented in this paper suggests that if household-level census data are not available, as is often the case, it is possible to generate reasonably accurate estimates of the incidence of poverty (P_0) using aggregated census data. The errors associated with aggregation are more likely to be acceptable if the level of aggregation of the census data is relatively low, such as at the district or enumeration area. Even highly aggregated census data can be used to rank provinces by poverty rate relatively accurately. The results in this paper provide information to help researchers anticipate the likely size and direction of the errors associated with using aggregate census data. In addition, researchers forced to work with aggregated census data can substantially reduce the aggregation errors if they can obtain from census authorities information on the variance in the estimated log per capita expenditure and apply the adjustment equation described in this paper.

Overall, these results suggest that, in some cases, high-resolution maps of the spatial patterns in poverty can be generated even in countries for which only aggregated census data are available. Such maps can contribute to efforts in these countries to alleviate poverty through geographically targeted policies and programs.

Appendix A: Derivation of error associated with using aggregate census data

This appendix derives an expression that describes the error associated with using aggregate census data instead of household-level census data in the second step of a poverty mapping analysis. We start with the second-order Taylor expansion:

$$f(x_1) \cong f(x_0) + (x_1 - x_0) f'(x_0) + \frac{1}{2}(x_1 - x_0)^2 f''(x_0)$$

If we duplicate this expression for N values of x, labelled $x_1 \ldots x_N$, and take the sum of the N equations, we get the following:

$$\sum_i f(x_i) \cong \sum_i f(x_0) + \sum_i (x_i - x_0) f'(x_0) + \frac{1}{2}\sum_i (x_i - x_0)^2 f''(x_0)$$

Dividing by N and setting the reference point (x_0) equal to the mean value of x ($\bar{x}$), the result is:

$$\frac{1}{N}\sum_i f(x_i) \cong f(\bar{x}) + \frac{1}{N}\sum_i (x_i - \bar{x}) f'(\bar{x}) + \frac{1}{2N}\sum_i (x_i - \bar{x})^2 f''(\bar{x})$$

But since the sum of deviations from the mean is zero, the second term on the right side drops out. Furthermore, the third term on the right side can be expressed in terms of the variance of x:

$$\frac{1}{N}\sum_i f(x_i) \cong f(\bar{x}) + \frac{1}{2}\text{var}(x_i) f''(\bar{x})$$

Next, we replace $f(.)$ with $\Phi(.)$, the cumulative standard normal distribution, and we replace x_i with $(\ln(z) - X_i^C \beta)/\sigma$, the difference between the log of the poverty line (z) and the estimated log per capita expenditure for household i ($X_i^C \beta$) divided by the standard error of the regression (σ). The result is:

$$\frac{1}{N}\sum_i \Phi\left[\frac{\ln(z) - X_i^C\beta}{\sigma}\right] \cong \Phi\left[\frac{1}{N}\sum_i \frac{\ln(z) - X_i^C\beta}{\sigma}\right] + \frac{1}{2}\text{var}\left(\frac{\ln(z) - X_i^C\beta}{\sigma}\right)\Phi''\left[\frac{1}{N}\sum_i \frac{\ln(z) - X_i^C\beta}{\sigma}\right]$$

If we assume that the poverty line (z) and the regression parameters (β and σ) are constant across the unit of aggregation of the census data, which will normally be the case,[13] then the first term on the right-hand side can be rewritten as follows:

$$\frac{1}{N}\sum_i \Phi\left[\frac{\ln(z) - X_i^C\beta}{\sigma}\right] \cong \Phi\left[\frac{\ln(z) - \bar{X}^C\beta}{\sigma}\right] + \frac{1}{2}\text{var}\left(\frac{\ln(z) - X_i^C\beta}{\sigma}\right)\Phi''\left[\frac{\ln(z) - \bar{X}^C\beta}{\sigma}\right]$$

[13] Typically, the regression analysis is carried out for urban and rural sectors or for each stratum of the household expenditure survey, so there are between 2 and 20 areas over which the regression parameters are constant. Similarly, the number of estimated poverty lines is usually relatively small (less than 20). By contrast, aggregated census data is often at the level of the district or enumeration area, of which there are generally more than 100. Thus, within a unit of aggregation, the poverty line and the regression parameters will, in most cases, be constant.

This equation describes the error associated with using aggregated census data instead of household-level data in estimating the proportion of *households* that are below the poverty line. If we wish to describe the errors in estimating the proportion of *people* below the poverty line, the averages in this equation must be rewritten as weighted averages, where the weights are the household size. This equation is further interpreted in Section 3.2.

References

Astrup, C. and S. Dessus (2001). 'The Geography of Poverty in the Palestinian Territories', *ERF Working Paper* 0120, Economic Research Forum: Cairo.

Bigman, D., S. Dercon, D. Guillaume, and M. Lambotte (2000). 'Community Targeting for Poverty Reduction in Burkina Faso', in D. Bigman and H. Fofack (eds), *Geographic Targeting for Poverty Alleviation: Methodology and Applications*, World Bank Regional and Sectoral Studies: Washington DC.

Chen, S., G. Datt and M. Ravallion (1991). 'POVCAL: A Program for Calculating Poverty Measures from Grouped Data' (mimeo), World Bank, Policy Research Department: Washington DC.

Elbers, C., J. Lanjouw and P. Lanjouw (2003). 'Micro-level Estimation of Poverty and Inequality', *Econometrica* 71:355-64.

Henninger, N. and M. Snel (2002). *Where are the Poor? Experiences with the Development and Use of Poverty Maps*, World Resources Institute: Washington DC, and UNEP/GRID: Arendal.

Hentschel, J., J. Lanjouw, P. Lanjouw and J. Poggi (1998). 'Combining Census and Survey Data to Study the Spatial Distribution of Poverty,' *World Bank Policy Research Working Papers* 1,928, World Bank: Washington DC.

Hentschel, J., J. Lanjouw, P. Lanjouw and J. Poggi (2000). 'Combining Census and Survey Data to Trace the Spatial Dimensions of Poverty: A Case Study of Ecuador', *World Bank Economic Review* 14:147-65.

Minot, N. (1998). 'Generating Disaggregated Poverty Maps: An Application to Vietnam', *Markets and Structural Studies Division Discussion Papers* 25, International Food Policy Research Institute: Washington DC.

Minot, N. (2000). 'Generating Disaggregated Poverty Maps: An Application to Vietnam,' *World Development* 28:319-31

Minot, N. and B. Baulch (2002). 'The Spatial Distribution of Poverty in Vietnam and the Potential for Targeting,' *World Bank Policy Research Working Papers* 2,829, World Bank: Washington DC.

Minot, N., B. Baulch, and M. Epprecht (2004). *Poverty and Inequality in Vietnam: Spatial Patterns and Geographic Determinants,* report prepared by the International Food

Policy Research Institute and the Institute of Development Studies in collaboration with the Inter-ministerial Poverty Mapping Working Group, Hanoi.

Poverty Working Group (1999). *Vietnam: Attacking Poverty*, a joint report of the Government of Vietnam-Donor-NGO Poverty Working Group presented to the Consultative Group Meeting for Vietnam, 14-15 December.

World Bank (1995). *Vietnam: Poverty Assessment and Strategy,* World Bank: Washington DC.

World Bank (2003). *World Development Indicators*, World Bank: Washington DC.