



*The World's Largest Open Access Agricultural & Applied Economics Digital Library*

**This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.**

**Help ensure our sustainability.**

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

[aesearch@umn.edu](mailto:aesearch@umn.edu)

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

*No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.*

**Poverty and Inequality in Vietnam**  
*Spatial Patterns and Geographic Determinants*

Nicholas Minot  
Bob Baulch  
Michael Epprecht

Copyright © 2006 International Food Policy Research Institute. All rights reserved.  
Sections of this material may be reproduced for personal and not-for-profit use  
without the express written permission of but with acknowledgment to IFPRI. To  
reproduce the material contained herein for profit or commercial use requires express  
written permission. To obtain permission, contact the Communications Division  
<ifpri-copyright@cgiar.org>.

International Food Policy Research Institute  
2033 K Street, N.W.  
Washington, D.C. 20006-1002  
U.S.A.  
Telephone +1-202-862-5600  
www.ifpri.org

DOI: 10.2499/0896291510

**Library of Congress Cataloging-in-Publication Data**

Minot, Nicholas.

Poverty and inequality in Vietnam : spatial patterns and geographic  
determinants / Nicholas Minot, Bob Baulch, and Michael Epprecht.

p. cm — (Research report / International Food Policy Research  
Institute ; 148)

Includes bibliographical references.

ISBN 0-89629-151-0 (alk. paper)

1. Poverty—Vietnam. 2. Equality—Vietnam. I. Baulch, Bob.  
II. Epprecht, Michael. III. Title. IV. Series: Research report  
(International Food Policy Research Institute) ; 148.

HC444.Z9P665 2006

305.5⋄6909597—dc22

2006010077

## Contents

List of Tables	iv
List of Figures	v
List of Boxes	vii
Foreword	viii
Summary	ix
1. Background	1
2. Data and Methods	4
3. Spatial Patterns in Poverty and Inequality	15
4. Geographic Determinants of Poverty	38
5. Spatial Variation in Determinants of Poverty	47
6. Summary and Conclusions	51
Appendix A: Comparison of Results Using Different Analysis Methods	58
Appendix B: Using GIS-Derived Variables for Statistical Analysis	65
References	71
<i>Color figures follow page 46</i>	

# Tables

1.1	Selected household data sets for Vietnam	2
2.1	Household characteristics in both the Census and the VLSS	6
2.2	Explanatory variables used in spatial regression analysis	12
3.1	Rural and urban regression models of per capita expenditure	17
3.2	Statistical significance of groups of variables	18
3.3	Comparison of poverty estimates at national and regional levels	19
3.4	Estimated poverty rate ( $P_0$ ) for urban and rural areas by province	20
3.5	Decomposition of inequality into between- and within-province components	31
3.6	Decomposition of inequality into between- and within-district components	31
4.1	Agro-climatic and socioeconomic factors that may affect poverty rate	39
4.2	Diagnostic tests for spatial dependence in rural poverty	40
4.3	Inclusive model of the geographic determinants of rural poverty	41
4.4	Selective model of the geographic determinants of rural poverty	43
4.5	Diagnostic tests for spatial dependence in urban poverty	44
4.6	Inclusive model of the geographic determinants of urban poverty	45
4.7	Selective model of the geographic determinants of urban poverty	46
5.1	Summary results of global model of rural poverty	48
5.2	Summary results of global and local models	48
5.3	Summary results of local parameter estimates	49
5.4	Significance of spatial variations in parameter estimates	50
A.1	Comparison of $P_0$ estimates using different methods	61
A.2	Comparison of Gini coefficient estimates using different methods	62
A.3	Comparison of poverty and inequality estimates using different methods	63
A.4	Correlation ( $R^2$ ) between poverty and inequality estimates	63

## Figures

1.1	Regions and provinces of Vietnam	Color insert
3.1	Map of the incidence of poverty ( $P_0$ ) for each province	Color insert
3.2	Provincial poverty rates and confidence intervals	22
3.3	Map of the incidence of poverty ( $P_0$ ) for each district	Color insert
3.4	District poverty rates and confidence intervals	24
3.5	Urban poverty and rural poverty by district	25
3.6	Map of the incidence of poverty ( $P_0$ ) for each commune	Color insert
3.7	Map of elevation and transportation infrastructure	Color insert
3.8	Commune poverty rates and confidence intervals	27
3.9	Map of the density of poverty	Color insert
3.10	Maps of the depth of poverty ( $P_1$ ) and severity of poverty ( $P_2$ ) for each district	Color insert
3.11	Depth of poverty ( $P_1$ ) and severity of poverty ( $P_2$ ) as a function of the incidence of poverty ( $P_0$ ) in each district	28
3.12	Map of inequality as measured by the Gini coefficient	Color insert
3.13	Maps of inequality as measured by the Theil L and Theil T indexes	Color insert
3.14	Theil indexes of inequality as a function of the Gini coefficient for each district	30
3.15	Poverty rate ( $P_0$ ) as a function of per capita expenditure	32
3.16	Gini coefficient of inequality as a function of per capita expenditure	33
3.17	Gini coefficient of inequality as a function of the poverty rate ( $P_0$ )	34
3.18	Poverty rate ( $P_0$ ) as a function of the share of the population in urban areas	34
3.19	Gini coefficient of inequality as a function of the share of the population in urban areas	35
3.20	Comparison of poverty rates ( $P_0$ ) from MOLISA and from small-area estimation methods	37
5.1	Maps of the spatial distribution of the independent variables	Color insert
5.2	Distribution of residuals in the global and local models	49

5.3	Map of the spatial distribution of local $R^2$	Color insert
5.4	Maps of the spatial distribution of the local coefficients of the independent variables	Color insert
B.1	Distance to nearest neighboring district measured from district centroids	70

## **Boxes**

3.1 Interpretation of the Standard Error and Confidence Interval	21
--	----

21



## Foreword

Information on the spatial distribution of poverty is particularly useful in designing geographically targeted programs to address regional disparities—a matter of high priority for many countries. Until recently, most developing countries were forced to design these programs based on rough indicators of poverty or on the results of household budget surveys, which typically generate poverty estimates for a limited number of regions.

In the late 1990s, a new approach was developed combining census data and household budget survey results to generate poverty estimates for small areas such as districts, allowing the construction of “poverty maps.” A 1998 IFPRI study in Vietnam was one of the first to experiment with this approach. A similar method was concurrently developed and subsequently refined by researchers at the World Bank. Poverty mapping studies have now been carried out in more than a dozen countries. This is an example of how an international public good can evolve from a set of specific country studies by the broader research community.

This report uses data from the 1998 Vietnam Living Standards Survey and the 1999 Population Census to estimate various measures of poverty and inequality for 614 districts in Vietnam. The results confirm conventional wisdom regarding high rates of poverty in the upland areas, low rates in urban areas, and intermediate rates in the irrigated lowland areas. But the results also offer some surprises, such as the fact that most poor people in Vietnam do not live in the poorest areas.

The study goes a step beyond standard poverty-mapping analysis by investigating the geographic determinants of district-level poverty rates. Distance to cities, soil type, and topography are significant predictors of local poverty rates. Surprisingly, proximity to small district centers is a stronger predictor of poverty than distance to large cities or distance to roads. Thus, rural-urban linkages at a local level matter a great deal for poverty reduction.

At the national level, this report provides valuable information for more precise geographic targeting of poverty assistance in Vietnam, and also offers insights into the geographic factors that contribute to rural poverty. Statistical authorities in Vietnam are now experimenting with variants of this method for use in poverty monitoring.

The study demonstrates the rewards that come from combining survey data, census data, and geographic data to focus on the challenge of reducing rural poverty. IFPRI continues to pursue this line of research, having recently carried out poverty-mapping analyses for Mozambique, Malawi, and Zambia.

Joachim von Braun  
Director General, IFPRI

## Summary

**T**his study uses a relatively new method called “small area estimation” to estimate various measures of poverty and inequality for provinces, districts, and communes of Vietnam. The method was applied by combining information from the 1997–98 Vietnam Living Standards Survey and the 1999 Population and Housing Census.

The results indicate that the poverty rate ( $P_0$ ) is greatest in the remote areas of the Northeast and Northwest, the upland areas of the North Central Coast, and the northern part of the Central Highlands. Poverty rates are intermediate in the Red River Delta and the Mekong River Delta. The lowest poverty rates are found in the main cities, Hanoi and Ho Chi Minh City, and in the Southeast region. The accuracy of these estimates is reasonable for the provincial and district estimates, but the commune estimates must be used with caution because some are not very precise.

Mapping the density of poverty reveals that, although the poverty rates are highest in the remote upland areas, these areas are sparsely populated, so most of the poor live in the Red River Delta and the Mekong River Delta.

Comparing these results with the district-level estimates of poverty from MOLISA, we find very little correlation. Several possible explanations for these differences are explored, but the most likely reason is variation in the methods used by MOLISA from one district to another.

This analysis confirms other studies indicating that the inequality in per capita expenditure is relatively low in Vietnam by international standards. Inequality is greatest in the large cities and (surprisingly) in parts of the upland areas. Inequality is lowest in the Red River Delta, followed by the Mekong Delta. Just one-third of the inequality is found between districts, and two-thirds within them, suggesting that district-level targeting of antipoverty programs may not be very effective.

District-level poverty is very closely associated with district-level average per capita expenditure. In other words, inequality does not explain much of the variation in poverty across districts.

We explored the geographic determinants of poverty using a global model (all rural areas) and a local model. In the global model, geographic determinants, including agro-climatic variables and market access, are able to explain about three-quarters of the variation in district-level rural poverty. Poverty is higher in districts with sloped land, bare and rocky land cover, soils that are poor (sandy, saline, or acid sulfate), and far from towns. By contrast, these agro-climatic and market access variables do not explain urban poverty very well.

The local regression model, in which coefficients vary from one area to another, reveals that flat land and high road density are associated with lower poverty throughout Vietnam. But other variables, such as rainfall and forest cover, are positively associated with poverty in some areas and negatively associated in others. Overall, the relationship between agro-climatic variables and poverty varies significantly from one area of Vietnam to another.

Many antipoverty programs are geographically targeted in Vietnam. The results from this study indicate that it may be possible to improve the targeting of these programs by adopting more precise estimates of poverty at the district and commune level, though further research

is needed to better understand the discrepancies between estimates produced by different methods.

The ability of market access and agro-climatic variables to explain a large portion of differences in rural poverty rates indicate that poverty in the remote areas is linked to low agricultural potential and lack of market access. This illustrates the importance of improving market access. The fact that poverty is closely related to low agricultural potential suggests that efforts to restrict migration out of disadvantaged regions may not be a good strategy for reducing rural poverty.

Finally, the study notes that the small-area estimation method is not very useful for annual poverty mapping because it relies on census data, but it could be used to show detailed spatial patterns in other variables of interest to policymakers, such as income diversification, agricultural market surplus, and vulnerability. Furthermore, it can be used to estimate poverty rates among vulnerable populations too small to be studied with household survey data, such as the disabled, small ethnic minorities, or fishermen.

## CHAPTER 1

---

### Background

In most countries, poverty is spatially concentrated. Extreme poverty in inaccessible areas with unfavorable soils and weather can often be found in the same country as relative affluence in more favorable locations close to major cities and markets. Information on the spatial distribution of poverty is of interest to policymakers and researchers for a number of reasons. First, it can be used to quantify suspected regional disparities in living standards and identify which areas are falling behind in the process of economic development. Second, it facilitates the targeting of programs, such as education, health, credit, and food aid, whose purpose is, at least in part, to alleviate poverty. Third, it may shed light on the geographic factors associated with poverty, such as mountainous terrain or distance from major cities.

In many countries, the main sources of information on spatial patterns of poverty are household income and expenditure surveys. These surveys generally have sample sizes of 2,000 to 8,000 households, which typically allow estimates of poverty for only 3 to 12 regions within a country. Research has shown that geographic targeting is most effective when the geographic units are quite small, such as a village or district (Baker and Grosh 1994; Bigman and Fofack 2000). The only household information usually available at this level of disaggregation is census data, but census questionnaires are generally limited to household characteristics and rarely include questions on income or expenditure.

In Vietnam there are at least three sources of information on the incidence of poverty. First, the General Statistics Office (GSO) has carried out two Vietnam Living Standards Surveys (VLSS), one in 1992–93 and the other in 1997–98. With samples of 4,800 and 6,000 households, respectively, these surveys generated poverty estimates for each of the seven regions of Vietnam.<sup>1</sup> Figure 1.1 (see color insert) provides the names and locations of these regions.

The GSO also carried out larger household surveys, such as the Multipurpose Household Survey and the 2001 Vietnam Household Living Standards Survey. These surveys have had samples of about 45,000 households and are intended to generate estimates that are valid at the provincial level. Figure 1.1 gives the names and locations of the 61 provinces in Vietnam,<sup>2</sup> and Table 1.1 summarizes the main sources of household survey data in Vietnam.

Another important source of information on the spatial distribution of poverty is the Ministry of Labor, Invalids, and Social Affairs (MOLISA). Each year, MOLISA prepares a list of

---

<sup>1</sup>Until 1998, the country was divided into seven regions for statistical purposes: the Northern Uplands (also called the North Mountains and Midland), the Red River Delta, the North Central Coast, the South Central Coast, the Central Highlands, the Southeast (also called the Northeast South), and the Mekong River Delta. In 1998, the Northern Uplands was split into the Northeast and the Northwest regions, but we retain the older grouping because the 1997–98 VLSS was designed to be representative at this level.

<sup>2</sup>In 2004, two provinces were split, making 63 provinces in total.

**Table 1.1 Selected household data sets for Vietnam**

Name of survey	Year	Sample (number of households)	Lowest level at which data are representative	Types of data collected	Use in this study
Vietnam Living Standards Survey	1992–93	4,800	Region	Income, expenditure, health, education, housing, assets, fertility, migration, etc.	Not used
Agricultural Census	1994	11,974,515	Any level	Land use, agriculture, housing, and assets	Not used
Multi-Purpose Household Survey	1994, 1995, 1996, 1997	45,000	Province	Income, expenditure, health, education, housing, assets, etc.	Not used
Vietnam Living Standards Survey	1997–98	5,999	Region	Income, expenditure, health, education, housing, assets, fertility, migration, etc.	Used in Stage 1 regression analysis
Population Census	1999	16,661,433	Any level	Household composition and housing	33% sample used for Stage 2 analysis
Vietnam Household Living Standards Survey	2002	75,000	Province	Income, expenditure, health, education, housing, assets, etc.	Not used

Sources: SDC/GSO (1994); GSO (1995, 1999, 2000).

poor households in each commune based on information gathered by local officials using criteria established by MOLISA. The welfare indicator is per capita income, where the poverty line is defined in terms of the value of a certain volume of rice at local prices. This information is used to identify the poorest communes, making them eligible for special programs and subsidies to reduce poverty. Although this system is relatively inexpensive and provides annual estimates, different provinces use somewhat different poverty lines and different data collection guidelines in implementing this analysis. Furthermore, even if the guidelines were made uniform, the use of thousands of enumerators to collect household-level data makes it difficult to ensure consistent application of those guidelines in the field (see Conway 2001).

In recent years, a new technique called small-area estimation has been developed that combines household and census data to estimate poverty rates (or other variables) for more disaggregated geographic units (see Elbers, Lanjouw, and Lanjouw 2003). Although various approaches have been used, they all involve three steps. First, one

selects household characteristics found in both the survey and the census, such as household composition, education, occupation, housing characteristics, and asset ownership. Second, the household survey data are used to generate an equation that estimates poverty or expenditure as a function of these household characteristics. Third, census data on those same household characteristics are inserted into the equation to generate estimates of poverty for small geographic areas.

For example, Minot (1998, 2000) used the 1992–93 Vietnam Living Standards Survey and a probit model to estimate the likelihood of poverty for rural households as a function of a series of household and farm characteristics. District-level means of these same characteristics were then obtained from the 1994 Agricultural Census and inserted into this equation, generating estimates of rural poverty for each of the 534 rural districts in the country (see Table 1.1 for more information on these data sources).

Elbers, Lanjouw, and Lanjouw (2003) developed a similar method, which was applied using survey and census data from Ecuador by Hentschel et al. (2000). By using

log-linear regression models and household-level data from a census, they were able to demonstrate that their method generates unbiased estimates of the headcount poverty rate and also to calculate the standard error of the estimated incidence of poverty.<sup>3</sup> This approach has been applied in at least a dozen countries, including Cambodia, Thailand, South Africa, and Panama (see Statistics South Africa and the World Bank 2000; World Bank 2000; Henninger and Snel 2002).

The earlier Vietnam study has several limitations. First, because it relied on the Agricultural Census, it generated poverty estimates only for the rural areas. Second, the use of a probit regression and district-level means, although intuitively plausible, does not generate unbiased estimates of district-level poverty (see Minot and Baulch 2002b for estimation of the size of this type of error). Third, in the absence of household-level census data, it was not possible to estimate the standard errors of the estimates to evaluate their accuracy.

More recently, Minot and Baulch (2002a) used the 1997–98 Vietnam Living Standards Survey and a 3 percent sample of the 1999 Population Census to generate estimates of the incidence of poverty in urban and rural areas of each of the 61 provinces in Vietnam. Unlike the earlier poverty-mapping analysis, this study uses household-level census data, allowing the calculation of the standard errors of the poverty estimates using the methods developed by Elbers, Lanjouw, and Lanjouw (2003) and Hentschel et al. (2000).

The present study has four objectives:

- To describe the spatial patterns in poverty and inequality in Vietnam
- To explore the geographic determinants (including agro-climatic factors and market access) of urban and rural poverty in Vietnam

- To examine the spatial variation in the relationship between poverty and the geographic determinants
- To draw implications from these results for the design of policies and programs in Vietnam and for further research.

This report expands on the previous study (Minot and Baulch 2002a) in three ways:

- By using a larger sample of the Census data (33 percent), it is able to provide estimates of the incidence of poverty for each district (of which there are 614) and for each commune (of which there are 10,747), although the latter are not very reliable. The earlier study calculated poverty estimates only for the 61 provinces.
- In addition to estimating the incidence of poverty, the current study also calculates two other poverty measures and three measures of inequality at the district level.
- Unlike the previous study, this analysis explores the geographic determinants (including agro-climatic factors and market access) of rural and urban poverty.

The report is organized in six sections. After this background section, Chapter 2 describes the data and methods used in this report. Chapter 3 examines the spatial patterns in poverty and inequality in Vietnam using three measures of poverty and three measures of inequality. Chapter 4 explores the geographic determinants of poverty, using spatial regression analysis and a set of variables extracted from geographic information systems (GIS) databases. Chapter 5 explores spatial variation in the relationship between poverty and the geographic factors using locally weighted regression. Finally, Chapter 6 summarizes the results and discusses some implications for policy and future research.

<sup>3</sup>The poverty headcount ratio is defined as the proportion of the population living in households with per capita expenditures below the poverty line.

## CHAPTER 2

---

### Data and Methods

#### Data

The poverty-mapping portion of this study makes use of two household data sets: the 1997–98 Vietnam Living Standards Survey (VLSS) and the 1999 Population and Housing Census. The VLSS was implemented by the General Statistics Office (GSO) of Vietnam with funding from the Swedish International Development Agency and the United Nations Development Program and with technical assistance from the World Bank. The sample includes 6,000 households in Vietnam, constituting a stratified random sample. The sample includes 4,270 households in rural areas and 1,730 households in urban areas. The quality of the VLSS survey data appears to be fairly good, judging by the level of effort in the design and implementation and by the small number of missing and out-of-range values (see GSO 2000).

The 1999 Population and Housing Census was carried out by the GSO and refers to the situation as of April 1, 1999. It was conducted with the financial and technical support of the United Nations Population Fund and the United Nations Development Program. The full results of the Census are not made available by the GSO, but we were able to obtain a 33 percent sample of the Census. The 33 percent sample was selected by GSO using systematic sampling of every third household on the list of households organized by administrative unit. The sample includes 5,553,811 households. Less information is available about the procedures used to collect the Census data, but the quality of the data appears to be good (Table 1.1 summarizes information on various sources of household data in Vietnam, including the 1997–98 VLSS and the 1999 Census).

The spatial analysis portion of this study used a variety of spatially referenced variables describing climate, topography, land cover, demographic, and market access. The topographic data were obtained from the United States Geological Survey (USGS). Data on roads and administrative boundaries were obtained from the Center for Remote Sensing and Geomatics, formerly attached to the General Department of Land Administration. Land cover, soil, and climate data were obtained from the Information Center for Agriculture and Rural Development of the Ministry of Agriculture and Rural Development. Finally, population and other demographic data were obtained from the 33 percent sample of the 1999 Population and Housing Census, described above. Many of these variables required considerable cleaning, processing, and further transformation in order to generate the variables used in the spatial analysis (see Appendix B for more information).

#### Methods to Estimate the Incidence of Poverty

The poverty line used in this study is the “overall poverty line” used in the analysis of the 1997–98 Vietnam Living Standards Survey (Poverty Working Group 1999; GSO 2000). The poverty line corresponds to the expenditure (including the value of home production and ad-

justed regional and seasonal price differences) required to purchase 2,100 kcal per person per day using the food basket of households in the third quintile plus a non-food allowance equal to what households in the third quintile spend on nonfood items. The poverty line was set at 1,789,871 VND/person per year, but the consumption expenditures in the survey were adjusted using monthly and regional price indexes to compensate for differences in the cost of living over the course of the survey and across regions.

Poverty mapping is one application of the method called small-area estimation. The method is typically divided into three stages:

- Stage 0 involves identifying variables that describe household characteristics that may be related to income and poverty and that exist in both the household survey and in the census.
- Stage 1 estimates a measure of welfare, usually per capita expenditure, as a function of these household characteristics using regression analysis and the household survey data.
- Stage 2 applies this regression equation to the same household characteristics in the census data, generating predicted welfare for each household in the census. This information is then aggregated up to the desired administrative unit, such as a district or province, to estimate the incidence of poverty and the standard error of the poverty estimate.

The three sections below describe these methods in more detail and describe how they were applied in the current study.

### **Stage 0: Identifying Household Characteristics in Both the VLSS and the Census**

The first step was to compare the questionnaires of the 1997–98 Vietnam Living Standards Survey and the 1999 Population and Housing Census to identify possible household characteristics found in both surveys that could be used as poverty indicators. In

addition to comparing the questionnaire, it is necessary to compare the values of the variables to ensure that they are in fact describing the same characteristics. For example, “type of employer” was initially considered for inclusion, but further investigation showed that a number of categories were defined differently in the VLSS and the Census, so the two could not be reconciled. As a result, this variable was excluded from the analysis. Based on this comparison, 17 household characteristics were selected for inclusion in the poverty mapping analysis (see Table 2.1).

Some household characteristics are categorical and, for regression analysis, must be represented by a number of dummy (binary) variables. For example, the main source of drinking water is a household characteristic, but for the regression analysis it must be represented by separate dummy variables for indoor tap, outdoor tap, covered well, uncovered well, and so on. Thus, the 17 household characteristics are represented in the regression analysis by 39 variables.

### **Stage 1: Estimating Per Capita Expenditure with a Household Survey**

As mentioned above, Stage 1 of the poverty mapping method involves using the household survey data and regression analysis to estimate household welfare as a function of household characteristics. In this study, we use real per capita consumption expenditure from the 1997–98 VLSS as the measure of household welfare. The explanatory variables are the 17 household characteristics described above, represented by 39 variables. Economic theory provides no guidance on the functional form, but generally a log-linear function is used:

$$\ln(y_i) = X_i' \beta + \varepsilon_i, \quad (1)$$

where  $y_i$  is the real per capita consumption expenditure of household  $i$ ,  $X_i'$  is a  $1 \times k$  vector of household characteristics of household  $i$ ,  $\beta$  is a  $k \times 1$  vector of estimated coefficients,



**Table 2.1 Household characteristics in both the Census and the VLSS**

Household characteristic	Question number	
	1999 Census	1997–98 VLSS
Household size (number of people)	Pt I, Q4	S1A
Proportion of household members over 60 years old	Pt I, Q4	S1A, Q2
Proportion of household members under 15 years old	Pt I, Q4	S1A, Q6
Proportion of household members who are women	Pt I, Q3	S1A, Q6
Highest level of education completed by head	Pt I, Q11–13	S2A
Whether or not the head of household has a spouse	Pt I, Q2	S1B, Q3
Highest level of education completed by spouse	Pt I, Q11–13	S2A
Whether or not head of household is an ethnic minority	Pt I, Q4	S0A
Occupation of head over last 12 months	Pt I, Q16	S4D
Type of house (permanent; semipermanent or wooden frame; “simple”)	Pt III, Q3	S6A, Q1
House type interacted with living area	Pt III, 4	S6C, Q1a
Whether or not household has electricity	Pt III, Q7	S6B, Q33
Main source of drinking water	Pt III, 8	S6B, Q25
Type of toilet	Pt III, Q9	S6B, Q31
Whether or not household owns a television	Pt III, Q10	S12C
Whether or not household owns a radio	Pt III, Q11	S12C
Region where household lives	Page 1	S0A

Source: Questionnaires for 1997–98 VLSS and 1999 Population and Housing Census.

and  $\varepsilon_i$  is a random disturbance term distributed as  $N(0, \sigma)$ . Because our main interest is predicting the value of  $\ln(y)$  rather than assessing the impact of each explanatory variable, we are not concerned about the possible endogeneity of some of the explanatory variables. Elbers, Lanjouw, and Lanjouw (2003) show that the probability that household  $i$  with characteristics  $X_i$  is poor can be expressed as:

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

where  $P_i$  is a variable taking a value of 1 if the household is poor and 0 otherwise,  $z$  is the “overall poverty line” (see GSO 2000, page 260), and  $\Phi$  is the cumulative standard normal function. If the predicted log per capita expenditure ( $X_i' \beta$ ) is equal to the log of the poverty line [ $\ln(z)$ ], then the term in brackets is zero, and the predicted probability that the household is poor is 50 percent. A lower predicted expenditure would imply a positive term in brackets and a higher prob-

ability that it is poor, whereas a higher predicted expenditure would imply a probability less than 50 percent.

### Stage 2: Applying Regression Results to the Census Data

In Stage 2 of the standard poverty-mapping method, the estimated regression coefficients from the first step are combined with census data on the same household characteristics to predict the probability that each household in the Census is poor. This is accomplished by inserting the household characteristics for household  $i$  from the census,  $X_i^C$ , into equation 2. The expected probability that household  $i$  is poor can be calculated as follows:

$$E[P_i | X_i^C, \beta, \sigma^2] = \Phi\left[\frac{\ln z - X_i^C \beta}{\sigma}\right]. \quad (3)$$

This estimate is not very accurate for an individual household, but it becomes more accurate when aggregated over many households. For a given area (such as a district or

province), Elbers, Lanjouw, and Lanjouw (2003) show that the proportion of the population living in households that are below the poverty line is estimated as the mean of the probabilities that individual households are poor:

$$E[P_i | X^C, \beta, \sigma^2] = \sum_{i=1}^N \frac{m_i}{M} \Phi \left[ \frac{\ln z - X_i^C \beta}{\sigma} \right], \quad (4)$$

where  $m_i$  is the size of household  $i$ ,  $M$  is the total population of the area in question,  $N$  is the number of households, and  $X$  is an  $N \times k$  matrix of household characteristics. The advantage of using the Census data, of course, is that the large number of households allows estimation of poverty headcount ratios for geographic units much smaller than would be possible with the VLSS data.

Provided that (1) the error term is homoskedastic, (2) there is no spatial autocorrelation, and (3) the full Census data are used, the variance of the estimated poverty headcount ratio can be calculated as follows:

$$\begin{aligned} \text{var}(P^*) &= \left( \frac{\partial P^*}{\partial \hat{\beta}} \right)' \text{var}(\hat{\beta}) \frac{\partial P^*}{\partial \hat{\beta}} \\ &+ \left( \frac{\partial P^*}{\partial \hat{\sigma}^2} \right)^2 \frac{2\hat{\sigma}^4}{n - k - 1} \\ &+ \sum_{i=1}^N \frac{m_i^2 P_i^* (1 - P_i^*)}{M^2}, \end{aligned} \quad (5)$$

where  $n$  is the sample size in the regression model. Thus,  $n$ ,  $k$ , and  $\sigma^2$  are from the regression analysis, whereas  $m_i$ ,  $M$ , and  $N$  are obtained from the census data. The partial derivatives of  $P^*$  with respect to the estimated parameters can be calculated as follows:

$$\frac{\partial P^*}{\partial \hat{\beta}_j} = \sum_{i=1}^N \frac{m_i}{M} \left( \frac{-x_{ij}}{\hat{\sigma}} \right) \phi \left( \frac{\ln z - X_i^C \hat{\beta}}{\hat{\sigma}} \right) \quad (6)$$

$$\frac{\partial P^*}{\partial \hat{\sigma}^2} = -\frac{1}{2} \sum_{i=1}^N \frac{m_i}{M} \left( \frac{\ln z - X_i^C \hat{\beta}}{\hat{\sigma}} \right) \phi \left( \frac{\ln z - X_i^C \hat{\beta}}{\hat{\sigma}} \right). \quad (7)$$

The first two terms in equation 5 represent the “model error,” which comes from the fact that there is some uncertainty regarding the true value of  $\beta$  and  $\sigma$  in the regression analysis. This uncertainty is measured by the estimated covariance matrix of  $\beta$  and the estimated variance of  $\sigma^2$  as well the effect of this variation on  $P^*$ . The third term in equation 5 measures the “idiosyncratic error,” which is related to the fact that, even if  $\beta$  and  $\sigma$  are measured exactly, household-specific factors will cause the actual expenditure to differ from predicted expenditure. These equations are described in more detail in Hentschel et al. (2000) and Elbers, Lanjouw, and Lanjouw (2003).

As noted above, equation 5 is valid only if the full Census data are available for the second stage of the mapping procedure. In this study, we use a 33 percent sample of the Census data in the second stage, so equation 5 must be modified as follows:

$$\begin{aligned} \text{var}(P^*) &= \left( \frac{\partial P^*}{\partial \hat{\beta}} \right)' \text{var}(\hat{\beta}) \frac{\partial P^*}{\partial \hat{\beta}} \\ &+ \left( \frac{\partial P^*}{\partial \hat{\sigma}^2} \right)^2 \frac{2\hat{\sigma}^4}{n - k - 1} \\ &+ \sum_{i=1}^N \frac{m_i^2 P_i^* (1 - P_i^*)}{M^2} + V_s, \end{aligned} \quad (8)$$

where  $V_s$  represents the variance associated with the sampling error in the Census, taking into account the design of the sample. In this study, we rely on the statistical software Stata to calculate the variance associated with the sampling error, taking into account the design of the sample.<sup>4</sup>

In order to compare poverty headcount ratios in different regions or provinces, it is

<sup>4</sup>This is accomplished with the “svymean” command. Stata calculates a linear approximation (a first-order Taylor expansion) of the sampling error variance based on information on the strata, the primary sampling unit, and the weighting factors. See StataCorp (2001) for more information.

convenient to calculate the variance of the difference between two estimates of poverty. Hentschel et al. (2000, footnote 17) provide an expression for the case when full Census data are used. Here we extend the expression to include the variance associated with sampling error:

$$\begin{aligned} \text{var}(P_1 - P_2) = & \left( \frac{\partial P_1 - P_2}{\partial \hat{\beta}} \right)' \text{var}(\hat{\beta}) \left( \frac{\partial P_1 - P_2}{\partial \hat{\beta}} \right) \\ & + \left( \frac{\partial P_1 - P_2}{\partial \hat{\sigma}^2} \right)^2 \frac{2\hat{\sigma}^4}{n - k - 1} \quad (9) \\ & + V_i(P_1) + V_i(P_2) + V_s(P_1) \\ & + V_s(P_2) - 2\text{cov}_s(P_1, P_2), \end{aligned}$$

where  $V_i(P_r)$  is the idiosyncratic variance of the poverty estimate for region  $r$  (the third term in equation 5),  $V_s(P_r)$  is the sampling variance of the poverty estimate for region  $r$ , and  $\text{cov}_s(P_1, P_2)$  is the covariance in the poverty estimates for regions 1 and 2 associated with sampling error.

## Methods to Estimate Other Measures of Poverty

The methods described above allow one to estimate the incidence of poverty, defined as the proportion of people below the poverty line. This measure, sometimes labeled  $P_0$ , is a member of a class of poverty measures identified by Foster, Greer, and Thorbecke (1984). These poverty measures can be expressed as follows:

$$P_\alpha = \frac{1}{N} \sum_{i=1}^M \left[ \frac{(z - y_i)}{z} \right]^\alpha, \quad (10)$$

where

- $z$  is the poverty line,
- $y_i$  is income (or expenditure) of person  $i$  in a poor household,
- $N$  is the number of people in the population, and

$M$  is the number of people in poor households.

Different values of  $\alpha$  in equation 10 give different poverty measures. When  $\alpha = 0$ , this formula gives the incidence of poverty. This is because the term in brackets is always 1, so the summation gives us the total number of people in poor households, which, when divided by  $N$ , gives us the proportion of people living in poor households. When  $\alpha = 1$ , it gives a measure called the depth of poverty (or the poverty gap).  $P_1$  takes into account not just how many people are poor but how poor they are on average. It is equal to the incidence of poverty ( $P_0$ ) multiplied by the average percentage gap between the poverty line and the income of the poor. When  $\alpha = 2$ , this equation gives a measure called the severity of poverty (or squared poverty gap).  $P_2$  takes into account not just how many people are poor and how poor they are but also the degree of income inequality among poor households. It is equal to the incidence of poverty ( $P_0$ ) multiplied by the average squared percentage gap between the poverty line and the income of the poor.

The poverty-mapping method described in the previous sections provides a method for estimating the proportion of people below a given poverty line,  $z$ , but do not provide any information on the distribution of income among the poor, which is necessary to calculate  $P_1$  and  $P_2$ . We can use the poverty-mapping method to estimate  $P_1$  and  $P_2$  by noting that  $z$  does not have to be the poverty line. We can estimate the cumulative distribution of the population by level of per capita expenditure by running the poverty-mapping calculations repeatedly for different values of  $z$ . More specifically, we use the following steps:

1. Select 100 levels<sup>5</sup> of per capita expenditure, divided evenly along the range

<sup>5</sup>The use of 100 levels is arbitrary. The larger the number of levels, the more accurate the estimation of the cumulative distribution and, hence, the more accurate the estimates of  $P_1$  and  $P_2$ . Of course, increasing the number of levels also increases the computational burden and time to run the program.

- of per capita expenditure from the poorest household to the richest household.
2. Set  $z$  equal to the lowest of these 100 levels (call this  $z_1$ ), and run the poverty-mapping calculations to calculate the proportion of the population with per capita expenditure below  $z_1$ .
  3. Then repeat step 2 setting  $z$  equal to each of the other 99 expenditure levels ( $z_2$  to  $z_{100}$ ), storing the values of  $z_i$  and the proportion of the population below  $z_i$  in a file for further analysis.

As  $z_i$  rises from its lowest level to its highest level, the proportion of people with per capita expenditure below  $z_i$  rises from 0 to 100 percent. Thus, these results trace out the cumulative distribution of the population by per capita expenditure.

This information can be used to calculate the values of  $P_1$  and  $P_2$ . In the gap between each pair of  $z$ 's ( $z_i$  and  $z_{i+1}$ ), we know the average per capita expenditure<sup>6</sup> and the proportion of people with per capita expenditures in that range. Thus, each pair of  $z$ 's that are below the poverty line can be used to represent one value of  $y_i$  in equation 10, taking into account the number of households with per capita expenditure in that range.

## Methods to Estimate Measures of Inequality

In this context, inequality measures describe the degree of variation in per capita expenditure across households. Perfect equality would describe the case in which all households have the same level of per capita expenditure, whereas perfect inequality would refer to the situation in which one household accounts for all the expenditure and others have none.

In this analysis, we calculate three of the more common measures of inequality: the

Gini coefficient, Theil's L index of inequality, and Theil's T index of inequality. The latter two measures are also part of a class of "general entropy" measures of inequality, so that the Theil L index is also called GE(0), and the Theil T index is also called GE(1).

The Gini coefficient is based on the Lorenz curve, which describes the cumulative distribution of income (or expenditure) as a function of the cumulative distribution of households. More specifically, the Gini coefficient is the area above the Lorenz curve and below the diagonal 45-degree line divided by the area under the diagonal line. When we have information about the proportion of people below different levels of per capita expenditure, the Gini coefficient can be approximated as follows:

$$\text{Gini} = 2 \sum_{i=1}^N \left[ \left( \frac{1}{2} (P_i + P_{i+1}) - \frac{1}{2} (X_i + X_{i+1}) \right) (P_{i+1} - P_i) \right], \quad (11)$$

where  $P_i$  is the cumulative share of the population for interval  $i$  and  $X_i$  is the cumulative share of expenditure for interval  $i$ . The first term in the large parentheses is the "height" of each slice, from the diagonal line down to the Lorenz curve, and the last term in small parentheses is the "width" of each slice. The Gini coefficient ranges from 0 (perfect equality) to 1 (perfect inequality).

The Theil L index of inequality is calculated as follows:

$$\text{Theil L} = \text{GE}(0) = \frac{1}{N} \sum_{i=1}^N \ln \left( \frac{\bar{y}}{y_i} \right), \quad (12)$$

where  $N$  is the number of households,  $\bar{y}$  is the average per capita expenditure, and  $y_i$  is the per capita expenditure of household  $i$ . The Theil L index ranges from 0 (perfect equality) to infinity (perfect inequality). This inequality measure gives greater weight to

<sup>6</sup>Strictly speaking, we know only the range of per capita expenditures in this group of households, and we assume that the average is  $(z_i + z_{i+1})/2$ . But if we choose a large number of  $z$ 's, the difference between  $z_i$  and  $z_{i+1}$  will be small, so the error in making this assumption will also be small.

the bottom end of the distribution. This implies that it gives greater weight to the distribution of expenditure among the poor than either the Gini coefficient or the Theil T index of inequality.

The Theil T index of inequality is calculated as:

$$\text{Theil T} = \text{GE}(1) = \frac{1}{N} \sum_{i=1}^N \frac{y_i}{\bar{y}} \ln\left(\frac{y_i}{\bar{y}}\right), \quad (13)$$

where the variables are defined as in equation 12. The Theil T index ranges from 0 (perfect equality) to  $\ln(N)$  (perfect inequality). This inequality measure gives equal weight to all parts of the distribution.

In order to calculate inequality measures, we use the three steps described above to generate the cumulative distribution of households by per capita expenditure. To estimate the Gini coefficient, we calculate the cumulative distribution of expenditure from the values of  $z_i$  and the cumulative proportion of the population from the values of  $P$  for each  $z_i$ . These can be used in equation 11 to calculate the Gini coefficient.

As in the calculation of  $P_1$  and  $P_2$ , the two Theil indexes of inequality are calculated by using each pair of  $z$ 's to represent one value of  $y_i$ . As described above, between each pair of  $z$ 's ( $z_i$  and  $z_{i+1}$ ), we know the average per capita expenditure and the proportion of people with per capita expenditure in that range. This information allows us to apply equations 12 and 13 to calculate the Theil indexes of inequality.

### Limitations of the Analysis

Three qualifications need to be made regarding the implementation of the poverty-mapping method in Vietnam. First, as in all poverty-mapping analyses, the requirement that all variables be in both the survey and the census constrains the number of variables that can be used to predict per capita expenditure. In particular, many of the explanatory variables are related to assets so they have a lagged relationship with per

capita expenditure. Ideally, it would be good to include variables that change with short-term fluctuations in per capita expenditure, such as food consumption patterns, but this information is not available in the census. However, as shown in the first section of this chapter, the explanatory power of our Stage 1 regression models is relatively good, providing some reassurance on this issue.

Second, the regression analysis in Stage 1 does not explicitly take into account heteroskedasticity (differences in the variance of the dependent variable across the sample). On the other hand, by expressing the dependent variable (per capita expenditure) as a logarithm, we reduce the degree of heteroskedasticity. In addition, we carry out the regression analysis with the “svyreg” command in Stata, which takes into account stratification and clustering in the sample in calculating the standard errors of the estimates. It does this by using the Huber/White/sandwich estimator of standard errors, which is robust to heteroskedasticity. The estimated coefficients are not biased, but they are “inefficient” in that they do not use all possible information (see StataCorp 2001, Volume 4, svyreg).

Third, the Stage 1 regression coefficients do not take into account spatial autocorrelation. Spatial autocorrelation exists when the dependent variable (or the error term) of the regression in a household in the VLSS is correlated with the dependent variable (or error term) in nearby households. If the error terms are correlated, the coefficients are unbiased but inefficient. This would be the case if some other factors (such as distance to a major city) were excluded from the regression model and spatially correlated. For example, all the households near a city might have negative error terms (predicted expenditure is less than actual expenditure). On the other hand, if the dependent variable in one household is directly affected by the value of a nearby household, then the estimated regression coefficients will be biased. One type of spatial autocorrelation is correlation among house-

holds in a sample cluster, sometimes called location effects (spatial autocorrelation is discussed in more detail in the next section of this chapter).

To reduce spatial autocorrelation, Elbers, Lanjouw, and Lanjouw (2003) recommend incorporating community-level variables in the Stage 1 regression model. These variables could be obtained by calculating community-level means of the household variables or by using geographic information systems (GIS) analysis to generate geographic variables representing climate, topography, or degree of market access. Although our analysis indicates the presence of some spatial autocorrelation, we were not able to eliminate it by including community-level variables in the regression analysis (more detail is given in Chapter 3). Furthermore, we were constrained from using geographic variables in the Stage 1 regression analysis because we plan to examine the geographic determinants of poverty at a later stage. We were concerned that including geographic variables in the Stage 1 model could exaggerate the strength of the relationship between (estimated) poverty and the geographic variables in the later analysis.

It should be noted that when heteroskedasticity and location effects are taken into account in the first-stage regressions, analytic solutions for the variance of the headcount ratio are not possible. Instead, it becomes necessary to use complex simulation methods to calculate the estimators and their standard errors (see Elbers, Lanjouw, and Lanjouw 2003). A program has been written in SAS software that uses the simulation approach proposed by Elbers, Lanjouw, and Lanjouw (2003) and takes heteroskedasticity and location effects into account. In Appendix A, we compare the results obtained from our Stata program with those obtained from the SAS program using a subsample of 26 districts in three provinces.

The results suggest that the district-level estimates of  $P_0$  and the Gini coefficient generated by the Stata program are reasonably accurate (unbiased, with an average error of less than 5 percent of the SAS value), though the standard errors of  $P_0$  may be underestimated. Furthermore, the Stata estimates of  $P_1$  and  $P_2$  are moderately accurate and highly correlated with the corresponding SAS estimates. However, the Stata estimates of the Theil L and Theil T indexes of inequality are less accurate, with an average error of around 20 percent of the SAS estimate. Appendix A provides more details on the results of this comparison.

## Methods in Global Spatial Regression Analysis

As discussed above, we are also interested in examining the geographic determinants of poverty. This analysis, sometimes called “Stage 3,” involves spatial regression analysis of poverty as a function of variables representing agro-climatic characteristics and market access. Because the dependent variable in this analysis is, itself, an imputed value, special care must be taken in interpreting the results, but Elbers, Lanjouw, and Lanjouw (2004) show that the basic results are essentially the same as they would be with a “true” measure of poverty.

This section describes the global spatial regression analysis, where “global” refers to the fact that the models assume that the relationship between poverty and geographic variables is the same across the country. The dependent variable is the district-level estimate of poverty obtained from Stage 2 of the poverty-mapping analysis described above. The independent variables are listed in Table 2.2.

As discussed, one of the problems with carrying out regression analysis on spatial relationships is that there is likely to be spatial autocorrelation in the data.<sup>7</sup> In general,

<sup>7</sup>A related problem is that geographic data must be aggregated to the level of administrative units, and the results may be affected by the way in which this is done. Appendix A describes this issue in more detail.

**Table 2.2 Explanatory variables used in spatial regression analysis**

Exogenous variables	Possibly endogenous variables
Percentage of area at different elevation ranges	Population and population density
Percentage of area at different slope ranges	Percentage of population in urban areas
Roughness of terrain	Number and types of markets
Soil type	Density of different types of roads
Type of land cover	Density of navigable rivers
Rainfall	Transport time to cities of different sizes
Temperature	
Hours of sunshine	
Distance to cities of different sizes	

spatial autocorrelation means that variables in one location are affected by the value of that variable in neighboring locations. There are two ways this problem can manifest itself.

Spatial lag dependence refers to a situation in which the dependent variable in one location is affected by the dependent variable in nearby locations. For example, if the dependent variable is income or poverty, it is probable that the level of economic activity in one location is directly affected by the level of economic activity in neighboring locations through migration, trade, or investment linkages. The spatial lag dependence model can be written as follows:

$$y_i = \sigma \sum_{j \neq i} w_{ij} y_j + X_i \beta + \varepsilon_i, \quad (14)$$

where

$y_i$  is the dependent variable for location  $i$ ,

$\sigma$  is the spatial autoregressive coefficient,

$w_{ij}$  is the spatial weight reflecting the proximity of  $i$  and  $j$ ,

$y_j$  is the dependent variable for location  $j$ ,

$X_i$  is a row vector of explanatory variables for location  $i$ ,

$\beta$  is a column vector of coefficients, and

$\varepsilon_i$  is the error term for location  $i$ .

The spatial weights matrix  $w$  describes the degree of proximity between each pair of spatial observations. Usually it is a binary variable based on whether the two locations

are contiguous or a continuous variable based on some function of the distance between the two locations. If the regression analysis is carried out without adjustment for spatial lag dependence, the estimated coefficients will be biased and inconsistent (Anselin 1988).

The second type of problem that may occur is spatial error dependence, in which the error term in one location is correlated with the error terms in nearby locations. This can occur if there are variables that are not included in the regression model but do have an effect on the dependent variable and they are spatially correlated. For example, the quality of local government affects income and poverty but is difficult to include in a regression model. Because the quality of local government is likely to be spatially correlated (all towns in a state are affected by the quality of state government), the error term in each location is likely to be correlated with those in nearby locations. This model can be written as follows:

$$y_i = X_i \beta + \varepsilon_i \text{ with } \varepsilon_i = \lambda \sum_{j \neq i} w_{ij} \varepsilon_j + u_i, \quad (15)$$

where

$y_i$  is the dependent variable for location  $i$ ,  
 $X_i$  is a row vector of explanatory variables for location  $i$ ,

$\beta$  is a column vector of coefficients,

$\varepsilon_i$  is the error term for location  $i$ ,

$\lambda$  is the spatial error autoregressive coefficient,

$w_{ij}$  is the spatial weight reflecting the proximity of  $i$  and  $j$ , and  
 $u_i$  is the uncorrelated portion of the error term for location  $i$ .

In this case, using ordinary least squares to estimate the model does not yield biased coefficients, but the estimates of the coefficient are not efficient, and the standard  $t$  and  $F$  tests will produce misleading inference (Anselin 1988).

In order to test for the presence of spatial autocorrelation, Moran's  $I$  is frequently used:

$$\text{Moran's } I = \frac{(x - \mu)'W(x - \mu)}{(x - \mu)'(x - \mu)}, \quad (16)$$

where

$x$  is a column vector of the variable of interest,  
 $\mu$  is the mean of  $x$ , and  
 $W$  is the weighting matrix.

This statistic is simply the correlation coefficient between  $x$  at one point in space and the weighted average of the values of  $x$  nearby. In order to test whether there is spatial lag dependence or spatial error dependence, the Lagrange multiplier is used to test the statistical significance of the spatial autocorrelation coefficient ( $\lambda$ ) in the two models. Anselin (1988) shows that the model with the larger coefficient ( $\lambda$ ) is likely to be the appropriate model.

In this study, we estimate the district level poverty rates ( $P_0$ ) as a function of the spatial variables listed in Table 2.1. A Chow test indicates that the coefficients to predict urban poverty differ significantly from the coefficients to predict rural poverty. Thus, we carry out the regression analysis separately for urban and rural areas.

The weighting matrix was generated using the inverse distance between the geographic centers of the two districts. In other words, the value of  $w_{ij}$  is equal to the inverse of the distance between the center of district  $i$  and the center of district  $j$ .

A Lagrange multiplier test is used to test for the statistical significance of  $\sigma$  and  $\lambda$ , which indicate the need to use the spatial dependence lag model or the spatial error dependence model, respectively. Often with spatial regression models, both parameters are statistically significant, and the normal procedure is to adopt the model that yields the higher value of the Lagrange multiplier.

The analysis was carried out using the "spatreg" module written for Stata and available from the Stata web site ([www.stata.com](http://www.stata.com)) as an add-on module.

## Methods in Local Spatial Regression Analysis

The global model described in the previous section assumes that the relationship between poverty and geographic factors is the same across the country. Local spatial regression analysis does not make this assumption and examines spatial variations in the relationship between poverty and geographic factors. We use a "moving window" regression framework in which numerous regression models are estimated, each centered on a "regression point" and including nearby observations defined by a fixed "kernel bandwidth." Coefficient estimates are generated for each regression point (see Fotheringham, Brunsdon, and Charlton 2002).

A model based on geographically weighted regression (GWR) techniques, where observations within the local regression window are weighted according to the distance to the regression point, was applied (Brunsdon, Fotheringham, and Charlton 1996). Observations closer to the regression point  $X_i$  receive more weight than data of observations further away. The weighted regression window is then "moved" to the next regression point, until all points have been covered.

Because this method is based on a conventional regression framework, the technique will produce the standard regression output for each regression point. This allows the regression output (including coefficients and  $R^2$ ) to be mapped, showing



their variation over space. This makes this technique particularly useful for analyzing relationships in spatial data.

The standard global regression model can be written as:

$$y_i = X_i' \beta + \varepsilon_i, \quad (17)$$

where

$y_i$  is the dependent variable,  
 $X_i'$  is a row vector of explanatory variables for location  $i$ ,  
 $\beta$  is a column vector of coefficients, and  
 $\varepsilon_i$  is the error term.

This model can be extended to a local regression model as follows:

$$y_i = X_i' \beta_i + \varepsilon_i, \quad (18)$$

where  $\beta_i$  is a column vector of coefficients specific to location  $i$ .

The regression coefficients vary from one observation to another because they are based on a local regression that includes observations in the vicinity. For each local regression at a regression point  $i$ , the observations are weighted depending on the distance from the regression point to the observation  $j$ . The Gaussian distance decay

function applied in this analysis can be written as follows:

$$w_{ij} = \exp\left[-\frac{1}{2} \frac{d_{ij}^2}{b}\right], \quad (19)$$

where

$w_{ij}$  is the weight at regression point  $i$  for observation  $j$ ,  
 $d_{ij}$  is the distance from regression point  $i$  to observation  $j$ , and  
 $b$  is the bandwidth or the radius of influence around each observation.

In addition, we can test whether a local model really describes relationships better than a global model by comparing global and local values of  $R^2$ . Furthermore, Fotheringham, Brunson, and Charlton (2002) proposed a Monte Carlo test of whether spatial variations in the estimated coefficients are statistically significant. The test involves randomly adjusting the geographic location of the observations numerous times, running a GWR on each, and then comparing statistically the parameter estimates for the randomly distributed observations with the parameter estimates of the actual geographic distribution.

## CHAPTER 3

---

### Spatial Patterns in Poverty and Inequality

#### Household Characteristics Correlated with Per Capita Expenditure

**A**s described above, the first step in constructing a poverty map is to estimate econometrically per capita consumption expenditure as a function of variables that are common to the Census and the VLSS. These household characteristics include household size and composition, ethnicity, education of the head of household and his or her spouse, occupation of the head of household, housing size and type, access to basic services, and ownership of selected consumer durables. Table 2.2 lists the variables used to represent these household characteristics in the regression analysis.

It is reasonable to expect that the coefficients to “predict” expenditure in rural areas may be different from those predicting expenditure in urban areas. Statistical tests indicate that the coefficients in the urban model are significantly different from those in the rural model.<sup>8</sup> This implies that separate analyses should be carried out on rural and urban samples.

In an earlier analysis, we tried estimating separate models for two urban and seven rural regions (see Minot and Baulch 2002a). The regression results were not very satisfactory, with lower values of  $R^2$ , more coefficients that were statistically insignificant, and some coefficients that had the “wrong” sign. Based on these results, we adopt the rural-urban regression models in this analysis.

We also experimented with the use of community-level means of the household-level variables as explanatory variables. This is recommended by Elbers, Lanjouw, and Lanjouw (2003) as a way to increase the explanatory power of the Stage 1 regression model and to reduce or eliminate spatial autocorrelation. Unfortunately, the community-level variables had very little effect on the explanatory power of the model and were not successful in reducing spatial autocorrelation.<sup>9</sup>

As discussed above, we decided not to use community-level geographic variables in Stage 1 because we wish to later examine the geographic determinants of the estimated poverty rates in Stage 3. The concern was that including geographic variables in Stage 1 could exaggerate the strength of the relationship between poverty and the geographic variables in Stage 3.

---

<sup>8</sup>The Chow test strongly rejects the hypothesis that the coefficients for the urban subsample are the same as those for the rural subsample ( $F = 6.16$ ,  $P < 0.001$ ).

<sup>9</sup>We added 18 commune-level means of variables representing household size, household composition, education, housing characteristics, electrification, type of toilet, type of water, and ownership of consumer durables. The additional variables increased the  $R^2$  but just 2 percentage points in rural and urban models. In addition, the spatial autocorrelation (as measured by the significance of commune-level dummy variables on the regression residuals) was still statistically significant in both models.

The results of the regression analysis are shown in Table 3.1. Both urban and rural models explain somewhat more than half of the variation in per capita expenditure. This is a relatively good result for cross-sectional data, but it is useful to keep in mind that other factors that are not included in the model explain almost half of the variation. In addition, because of the endogeneity of some of the “explanatory” variables, the statistical significance of the coefficients should not be interpreted as implying causality, nor should much weight be given to the size of the coefficients.

According to the results in Table 3.1, large households are strongly associated with lower per capita expenditure in both urban and rural areas. Given the likely economies of scale in family size, these results do not necessarily imply a negative relationship between welfare and household size.

In rural areas, a household with a large proportion of elderly members, of children, and of women is likely to be poorer, other factors being equal. In urban areas, however, only the share of children is associated with poverty. This implies that income-earning capacity in urban areas is less dependent on physical strength than it is in rural areas.

Ethnicity<sup>10</sup> is a surprisingly weak predictor of per capita expenditure after other household characteristics are controlled. In rural areas the coefficient on ethnicity was significant only at the 10 percent level, whereas in urban areas it was not statistically significant. In both urban and rural areas the level of schooling of the head of household is a good predictor of a household’s per capita expenditure (the omitted category is no schooling). The five variables that represent the education of the head of

household are jointly significant at the 1 percent level in both rural and urban areas (see Table 3.1).

In general, the educational level of the spouse is less significant than that of the household head as a predictor of per capita expenditure.

The occupation of the head of household is a statistically significant predictor of per capita expenditure in rural and urban areas, other factors held equal. As expected, a household whose head is working in a skilled occupation is better off than other households.

Housing characteristics are good predictors of expenditures. Living in a house made of permanent or semipermanent materials is associated with significantly higher per capita expenditure in both rural and urban areas.<sup>11</sup> The living area of a house is also a useful predictor of household well-being (houses in Vietnam have an average living area of about 45 square meters). Electrification<sup>12</sup> is a statistically significant predictor of household welfare in rural areas, where 71 percent of the households have access to electricity, but not in urban areas, where 98 percent of the households are already electrified (see Table 3.1).

The main source of water and type of sanitation facilities are also useful predictors. In rural areas, households with access to well water have a higher level of per capita expenditures than households using river or lake water (the omitted category). In urban areas, where more than half of the sample households (58 percent) have access to tap water, this variable is a good predictor of urban per capita expenditures.

In rural areas, flush toilets and latrines are statistically significant indicators of

<sup>10</sup>Ethnic minorities are defined as all ethnic groups except for Kinh (ethnic Vietnamese) and Hoa (ethnic Chinese), following the classification commonly used in Vietnam.

<sup>11</sup>The effect of permanent and semipermanent housing materials operates through both the dummy variables for these characteristics and the interaction terms with living areas. Although the dummy variables have negative coefficients, this is more than offset by the positive effect through the interaction terms.

<sup>12</sup>More specifically, this variable refers to the main type of lighting used by the households.

**Table 3.1 Rural and urban regression models of per capita expenditure**

Variable	Rural model <sup>a</sup>		Urban model <sup>b</sup>	
	Coefficient	<i>t</i>	Coefficient	<i>t</i>
Size of household (members)	-0.0772	-19.5***	-0.0785	-8.1***
Proportion over 60 years (fraction)	-0.0831	-2.4**	-0.1026	-1.6
Proportion under 15 years (fraction)	-0.3353	-9.4***	-0.2368	-3.6***
Proportion female (fraction)	-0.1177	-3.5***	0.0386	0.5
Household head is ethnic minority	-0.0765	-1.9*	0.0142	0.2
Head has completed primary school	0.0585	3.4***	0.0616	1.7
Head has completed lower secondary school	0.0883	4.5***	0.0338	1.3
Head has completed upper secondary school	0.0884	3.3***	0.1368	3.2***
Head has completed advanced technical degree	0.1355	4.2***	0.1603	3.5***
Head has postsecondary education	0.2552	4.9***	0.1843	3.7***
Head does not have a spouse	0.0173	1.0	0.0344	0.8
Spouse has completed primary school	0.0049	0.3	0.0642	1.9*
Spouse has completed lower secondary school	0.0132	0.6	0.0987	2.6**
Spouse has completed upper secondary school	0.0107	0.3	0.1912	2.7**
Spouse has completed advanced technical degree	0.0921	2.3**	0.1285	3.2***
Spouse has postsecondary education	0.1571	2.7***	0.1752	3.1***
Head is a political leader or manager	0.1414	3.5***	0.2312	3.0***
Head is a professional or technical worker	0.1350	3.3***	0.0576	1.2
Head is a clerk or service worker	0.1362	3.4***	0.0357	0.9
Head is in agriculture, forestry, or fishing	-0.0163	-0.6	-0.0093	-0.2
Head is a skilled worker	0.0701	1.9*	0.0071	0.2
Head is an unskilled worker	-0.0586	-1.7*	-0.1599	-2.9***
House made of permanent materials	-0.9228	-4.3***	-0.5194	-3.4***
House made of semipermanent materials	-0.3120	-3.6***	-0.4001	-3.8***
Interaction of log(house area) and permanent house	0.2958	5.7***	0.2001	5.4***
Interaction of log(house area) and semipermanent house	0.1180	5.2***	0.1403	4.6***
House has electricity	0.0765	2.7***	-0.0026	0.0
House uses water from a public or private tap	0.0828	1.4	0.2289	5.3***
House uses well water	0.1157	4.4***	0.0340	0.6
House has flush toilet	0.2700	5.5***	0.1311	2.2**
House has latrine	0.0556	2.6**	0.0049	0.1
Household has television	0.2124	15.1***	0.2167	5.5***
Household has 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**	omitted	
Southeast	0.5424	9.4***	0.4151	5.5***
Mekong River Delta	0.3011	5.1***	0.1895	2.1**
Constant	7.5327	108.7***	7.7538	64.7***

Source: Regression analysis of 1997–98 Vietnam Living Standards Survey, taking into account clustering and stratification and using robust estimates of standard errors.

Notes: Omitted categories are: head has no education; spouse has no education; head is not working; house is made of temporary materials; household has other water source; household has no sanitation facilities; and household lives in the Northern Uplands. \* indicates significance at the 10 percent level, \*\* at the 5 percent level, and \*\*\* at the 1 percent level.

<sup>a</sup>*N* = 4,269, *R*<sup>2</sup> = 0.536.

<sup>b</sup>*N* = 1,730, *R*<sup>2</sup> = 0.550.

**Table 3.2 Statistical significance of groups of variables**

Sector	Variable	df <sub>1</sub>	df <sub>2</sub>	F statistic	Probability
Rural	Education of head of household	5	129	7.80	0.0000***
	Education of spouse	6	129	1.97	0.0738*
	Occupation of head of household	6	129	12.65	0.0000***
	Type of housing	2	129	14.00	0.0000***
	Main source of water	2	129	9.69	0.0001***
	Type of sanitary facility	2	129	15.64	0.0000***
	Region	6	129	26.20	0.0000***
Urban	Education of head of household	5	55	4.01	0.0036***
	Education of spouse	6	55	3.10	0.0110**
	Occupation of head of household	6	55	2.90	0.0157**
	Type of housing	2	55	10.76	0.0001***
	Main source of water	2	55	17.17	0.0000***
	Type of sanitary facility	2	55	4.12	0.0216**
	Region	5	55	10.29	0.0000***

Source: Regression analysis of per capita expenditure using 1997–98 VLSS.

Notes: The dependent variable is log of per capita expenditure. \* indicates significance at the 10 percent level, \*\* at the 5 percent level, and \*\*\* at the 1 percent level.

higher per capita expenditure, whereas in urban areas, having a flush toilet is a statistically significant predictor (see Table 3.1).

Television and radio ownership are two of the strongest predictors of per capita expenditures. Both variables are statistically significant in both urban and rural areas. As expected, the coefficient for radio ownership is smaller than that of television ownership.

Regional dummy variables were included in the regression models, with the Northern Uplands as the omitted region. Even after controlling for other household characteristics, rural households in the four southern regions are shown to be better off than those in the Northern Uplands. A similar pattern holds for urban households (see Table 3.1). The regional dummy variables are jointly significant at the 1 percent level in both urban and rural areas (see Table 3.2).

### Incidence of Poverty

The incidence of poverty (also called the poverty rate or poverty headcount ratio) is defined here as the proportion of the population living in households whose per capita expenditure is below the “overall poverty line,” as defined by the GSO (2000, 260).

This is the Foster-Greer-Thorbecke measure of poverty when  $\alpha = 0$ , also known as  $P_0$ . We present national, regional, provincial, district, and commune-level estimates of the poverty rate in turn.

### National and Regional Poverty Rates ( $P_0$ )

The national headcount poverty rate, as estimated in this application of the small-area estimation method using a 33 percent sample of the 1999 Census data, is 36.5 percent, less than 1 percentage point from the estimate from the 1997–98 VLSS (see Table 3.3). The small-area estimate of the urban poverty rate (11.6 percent) is 2.4 percentage points lower than the corresponding estimate from the 1997–98 VLSS, but the small-area estimate for the rural poverty rate (44.3 percent) is 1.2 percentage points higher.

The difference in regional poverty estimates ranges from less than 2 percentage points in the North Central Coast, Southeast, and Mekong River Delta to 8.9 percentage points in the Central Highlands. It should be noted that the 1997–98 VLSS had a relatively small sample for the Central Highlands, just 368 households, and that there were no urban households sampled in this

**Table 3.3 Comparison of poverty estimates at national and regional levels**

Level	Headcount poverty rate (percent)		Difference (percentage points)
	1997–98 VLSS	Small-area estimation method	
National	37.4	36.5	0.9
By urban/rural residence			
Urban	9.2	11.6	–2.4
Rural	45.5	44.3	1.2
By region			
Northern Uplands	58.6	53.6	5.0
Red River Delta	28.7	31.6	–2.9
North Central Coast	48.1	46.2	1.9
South Central Coast	35.2	39.1	–3.9
Central Highlands	52.4	43.5	8.9
Southeast	7.6	8.5	–0.9
Mekong River Delta	36.9	35.7	1.2

region. In contrast, the small-area estimates are based on 273,035 Census households in the Central Highlands, 27 percent of which live in urban areas. In fact, the estimate for rural poverty in the Central Highlands using small-area estimation is 53.7 percent, just 1.3 percentage points above the VLSS estimate. This suggests that the small-area estimates may be more accurate than survey-based estimates in some cases.

### Provincial Poverty Rates ( $P_0$ )

Table 3.4 gives the estimates of the incidence of poverty for each of the 61 provinces in Vietnam, and Figure 3.1 (see color insert) maps these estimates. In the map, the poorest areas are dark orange, and the least poor areas are dark green. These results confirm that poverty is most widespread in the Northwest and the Northeast, particularly in the provinces along the northern border with China and the northwestern border with the Lao P.D.R. More specifically, the poverty rate is highest (70–80 percent) in the provinces colored dark orange: Lai Chau, Ha Giang, and Son La. The light orange indicates that the poverty rate is 60–70 percent in Lao Cai, Cao Bang, Lang Son, and Bac Kan. Poverty is lower but still above 50 percent in the yellow provinces, including the interior provinces of the Northeast and

Northwest (Hoa Binh, Tuyen Quang, and Yen Bai), the northern part of the Central Highlands (Gia Lai and Kon Tum), and two central coast provinces (Ninh Thuan and Quang Tri). It is striking that the 10 poorest provinces are all in the Northeast and Northwest.

The provinces where the poverty rate is lowest are, not surprisingly, the ones near large urban centers. Three southern provinces (Ho Chi Minh, Binh Duong, and Ba Ria-Vung Tau) are dark green, implying that they have estimated poverty rates around 10 percent or lower. Other provinces, including Dong Nai, Tay Ninh, Da Nang, Hanoi, and Binh Phuoc, have estimated poverty rates between 10 and 20 percent.

Most provinces in the Red River Delta, the Mekong Delta, and the Central Coast are various shades of lighter green, indicating a poverty rate in the range of 20–50 percent.

Although the poverty map is useful for identifying the spatial patterns of poverty, Table 3.4 provides more detail, including the standard errors of the poverty estimates and the urban and rural poverty rates for each province. One of the strengths of this poverty-mapping method is that it calculates the standard errors, a measure of the accuracy of the estimate (see Box 3.1 for more explanation of standard errors).

Table 3.4 Estimated poverty rate ( $P_0$ ) for urban and rural areas by province

Code	Province	Rank (1 = poorest)	Overall		Rural		Urban	
			Poverty rate ( $P_0$ )	Standard error	Poverty rate ( $P_0$ )	Standard error	Poverty rate ( $P_0$ )	Standard error
101	Ha Noi	55	0.16	0.013	0.31	0.026	0.05	0.011
103	Hai Phong	51	0.29	0.020	0.40	0.029	0.08	0.017
105	Ha Tay	33	0.39	0.027	0.41	0.029	0.13	0.024
107	Hai Duong	47	0.33	0.026	0.36	0.030	0.12	0.026
109	Hung Yen	39	0.37	0.027	0.39	0.029	0.17	0.036
111	Ha Nam	35	0.38	0.028	0.40	0.030	0.14	0.028
113	Nam Dinh	41	0.35	0.026	0.38	0.030	0.11	0.024
115	Thai Binh	43	0.34	0.029	0.36	0.030	0.08	0.020
117	Ninh Binh	35	0.38	0.025	0.42	0.029	0.10	0.023
201	Ha Giang	2	0.75	0.020	0.81	0.022	0.24	0.033
203	Cao Bang	5	0.67	0.024	0.75	0.027	0.17	0.041
205	Lao Cai	4	0.70	0.018	0.79	0.021	0.21	0.027
207	Bac Kan	7	0.60	0.027	0.67	0.031	0.21	0.037
209	Lang Son	6	0.62	0.024	0.73	0.028	0.17	0.033
211	Tuyen Quang	9	0.57	0.030	0.61	0.033	0.14	0.023
213	Yen Bai	9	0.57	0.025	0.67	0.031	0.17	0.025
215	Thai Nguyen	25	0.43	0.033	0.50	0.042	0.16	0.022
217	Phu Tho	20	0.45	0.038	0.50	0.044	0.15	0.022
219	Vinh Phuc	20	0.45	0.044	0.48	0.049	0.20	0.028
221	Bac Giang	17	0.46	0.042	0.48	0.046	0.16	0.024
223	Bac Ninh	35	0.38	0.043	0.40	0.048	0.18	0.026
225	Quang Ninh	41	0.35	0.025	0.51	0.041	0.15	0.024
301	Lai Chau	1	0.80	0.014	0.88	0.015	0.19	0.029
303	Son La	3	0.73	0.020	0.81	0.022	0.14	0.023
305	Hoa Binh	8	0.59	0.028	0.65	0.032	0.14	0.023
401	Thanh Hoa	17	0.46	0.034	0.49	0.037	0.14	0.025
403	Nghe An	17	0.46	0.034	0.50	0.037	0.12	0.022
405	Ha Tinh	20	0.45	0.036	0.48	0.040	0.15	0.028
407	Quang Binh	15	0.47	0.033	0.51	0.038	0.14	0.026
409	Quang Tri	13	0.51	0.027	0.59	0.034	0.22	0.031
411	Thua Thien-Hue	15	0.47	0.026	0.58	0.036	0.20	0.027
501	Da Nang	55	0.16	0.017	0.34	0.032	0.11	0.020
503	Quang Nam	29	0.41	0.029	0.46	0.034	0.17	0.029
505	Quang Ngai	20	0.45	0.030	0.49	0.034	0.15	0.027
507	Binh Dinh	35	0.38	0.028	0.45	0.036	0.16	0.028
509	Phu Yen	29	0.41	0.029	0.46	0.035	0.18	0.031
511	Khanh Hoa	47	0.33	0.022	0.44	0.032	0.14	0.022
601	Kon Tum	13	0.51	0.037	0.65	0.052	0.20	0.028
603	Gia Lai	11	0.53	0.037	0.63	0.049	0.20	0.027
605	Dak Lak	25	0.43	0.045	0.50	0.056	0.17	0.025
701	TP Ho Chi Minh	61	0.05	0.008	0.08	0.012	0.05	0.009
703	Lam Dong	43	0.34	0.035	0.46	0.055	0.14	0.021
705	Ninh Thuan	11	0.53	0.026	0.63	0.032	0.21	0.032
707	Binh Phuoc	54	0.17	0.020	0.19	0.023	0.08	0.017
709	Tay Ninh	57	0.13	0.016	0.14	0.018	0.08	0.017
711	Binh Duong	60	0.08	0.008	0.08	0.011	0.06	0.012
713	Dong Nai	58	0.11	0.012	0.13	0.016	0.06	0.010
715	Binh Thuan	20	0.45	0.026	0.53	0.034	0.24	0.035
717	Ba Ria-Vung Tau	59	0.10	0.010	0.13	0.016	0.06	0.010
801	Long An	51	0.29	0.022	0.32	0.026	0.14	0.026

Table 3.4—Continued

Code	Province	Rank (1 = poorest)	Overall		Rural		Urban	
			Poverty rate ( $P_0$ )	Standard error	Poverty rate ( $P_0$ )	Standard error	Poverty rate ( $P_0$ )	Standard error
803	Dong Thap	33	0.39	0.025	0.42	0.028	0.19	0.032
805	An Giang	31	0.40	0.022	0.46	0.027	0.21	0.032
807	Tien Giang	53	0.27	0.024	0.30	0.027	0.12	0.024
809	Vinh Long	47	0.33	0.024	0.36	0.027	0.15	0.027
811	Ben Tre	50	0.32	0.024	0.34	0.026	0.15	0.028
813	Kien Giang	31	0.40	0.023	0.45	0.027	0.21	0.034
815	Can Tho	43	0.34	0.022	0.40	0.027	0.14	0.026
817	Tra Vinh	25	0.43	0.026	0.47	0.030	0.19	0.030
819	Soc Trang	25	0.43	0.025	0.48	0.029	0.22	0.033
821	Bac Lieu	40	0.36	0.023	0.41	0.028	0.20	0.034
823	Ca Mau	43	0.34	0.024	0.38	0.028	0.17	0.028

Source: Analysis of 1997–98 VLSS and 1999 Population and Housing Census.

Notes: The poverty rate refers to the proportion of the population that are in households whose per capita expenditure is below the overall poverty line. The standard error is a measure of the accuracy of the poverty estimate. The 95 percent confidence interval is approximately  $\pm 2$  times the standard error.

The poverty rates and their respective confidence intervals for the 61 provinces (in order from least poor to most poor) are shown in Figure 3.2. The diamond-shaped markers are the provincial poverty estimates,

and the horizontal lines above and below each estimate are the upper and lower limits of its 95 percent confidence interval. This graph shows that most of the provinces (36) have poverty rates in the range of 20

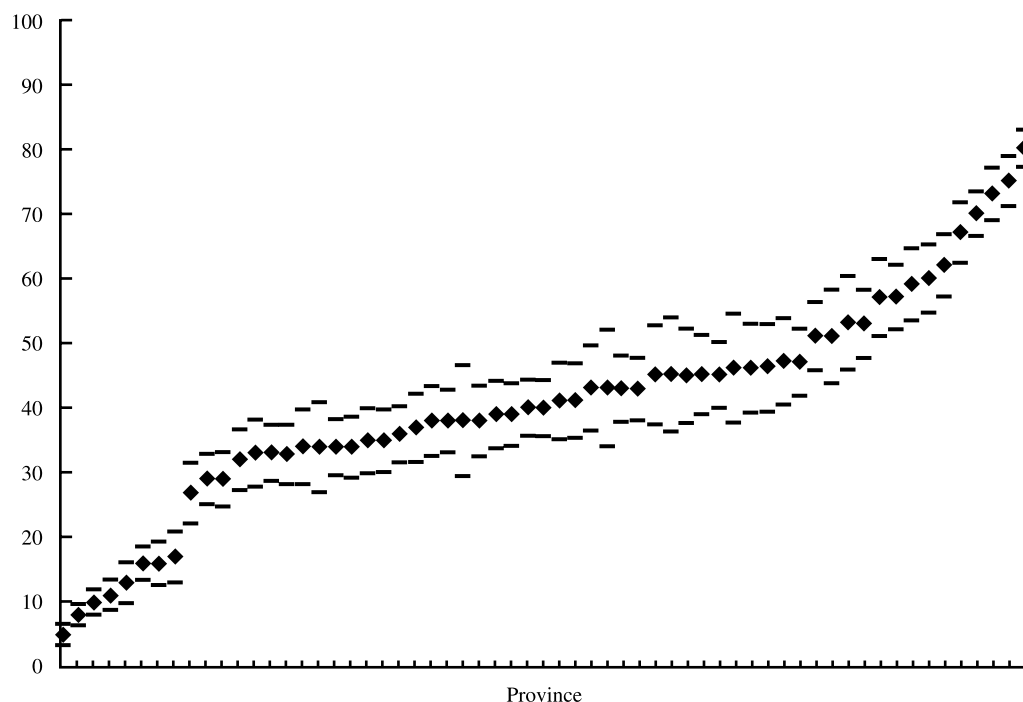
### Box 3.1 Interpretation of the Standard Error and Confidence Interval

Like any method for measuring poverty, the poverty-mapping method does not produce exact results. The household characteristics do not perfectly predict household expenditure in Stage 1. Even if they did, there may be differences between households in the VLSS sample and those in the Census. Finally, our Census data consist of a 33 percent sample of the original data, so there is some sampling error as well.

A number of factors affect the accuracy of the poverty estimate. First, if the Stage 1 regression equation is very good in predicting household expenditure based on the household characteristics, then the poverty estimates will be more accurate. Second, the accuracy of poverty estimates tends to be better for areas with poverty rates near 0 percent or near 100 percent. Third, the accuracy is better for areas with a large number of similar households than for areas with few and diverse households.

Standard errors help define the margin of error around the poverty estimates. There is a 95 percent chance that the “true” poverty estimate lies within two standard errors of the poverty estimate. For example, in the case of Yen Bai, the estimated poverty rate is 0.57 (57 percent), and the standard error is 0.025. This means the 95 percent confidence interval of this poverty estimate is 57 percent  $\pm$  5 percentage points ( $0.025 \times 2$ ). In other words, there is a 95 percent chance that the true poverty rate for Yen Bai is between 52 and 62 percent.



**Figure 3.2 Provincial poverty rates and confidence intervals**Poverty rate ( $P_0$ ) and confidence interval (%)

to 50 percent. In contrast, there are fewer provinces with poverty rates below 20 percent or above 50 percent, and the gap between the poverty rates of adjacent provinces is relatively large.

Across provinces, the 95 percent confidence interval (see Box 3.1) ranges from  $\pm 1.6$  percentage points to  $\pm 9$  percentage points, with the average confidence interval being  $\pm 5.2$ . Half the provinces have confidence intervals between  $\pm 4.2$  and  $\pm 5.8$  percentage points.<sup>13</sup> Dak Lak has the highest confidence interval ( $\pm 9$  percentage points). It is not obvious why the standard error is so large for Dak Lak; it may be that households earning money from coffee spend it in ways that are not captured by our 17 household characteristics.

One important implication of these standard errors is that if two provinces have poverty rates that differ by 5 percentage points, for example, there is a good chance that the difference is not statistically significant. For example, if province A has a poverty rate of 40 percent and province B has a rate of 44 percent, we generally cannot say that province B is poorer than province A. As a general rule, two poverty rates must differ by at least 8–10 percentage points to give us confidence that the difference is statistically significant.

As can be seen in Figure 3.2, the confidence intervals are much smaller when the poverty rate is either quite high or quite low. When the poverty rate is below 20 percent or above 70 percent, the confidence interval

<sup>13</sup>More specifically, this refers to the inter-quartile range. The first figure ( $\pm 4.2$ ) is the 25th percentile and the second ( $\pm 5.8$ ) is the 75th percentile. Thus, half the provinces lie within this range.

tends to be  $\pm 2$  or  $\pm 3$  percent. In contrast, when the poverty rate is around 50 percent, the confidence interval tends to be  $\pm 5$  percentage points or more.

It is also useful to look at differences in the incidence of poverty in urban and rural areas within each province (see Table 3.4). In all 61 provinces, the rural poverty rate is higher than the urban poverty rate. In fact, whereas the rural poverty rate ranges widely from 8 percent to almost 90 percent, the urban poverty rates are all less than 25 percent.

### District Poverty Rates ( $P_0$ )

The poverty-mapping method can also be used to generate poverty estimates for each of the 614 districts in Vietnam. The spatial patterns in the incidence of poverty can be seen in Figure 3.3 (see color insert). The district-level poverty map shows considerably more detail than the provincial poverty map. For example, in the provincial map, Son La and Lai Chau appear as one orange block, implying a poverty rate in the range of 70–80 percent. The district map, however, shows that the poverty rate varies widely within these two provinces, being under 70 percent in southeastern Son La (Yen Chau, Phu Yen, and Moc Chau districts) and over 90 percent in the far northwestern corner of Lai Chau province (Mong Te and Sin Ho districts).

Similarly, the provincial map suggests that almost all of the North Central Coast is light green, implying poverty rates in the 40–50 percent range. In contrast, the district map shows that the poverty rate along the coastal plain is less than 40 percent, but the rates in the interior are greater than 70 percent for some districts. Two districts in this region have poverty rates over 80 percent, both on the Lao border: Muong Lat district in Thanh Hoa and Ky Son district in Nghe An. The district map also reveals variation in the incidence of poverty in the Central Highlands and in the Mekong Delta that are hidden in the provincial map.

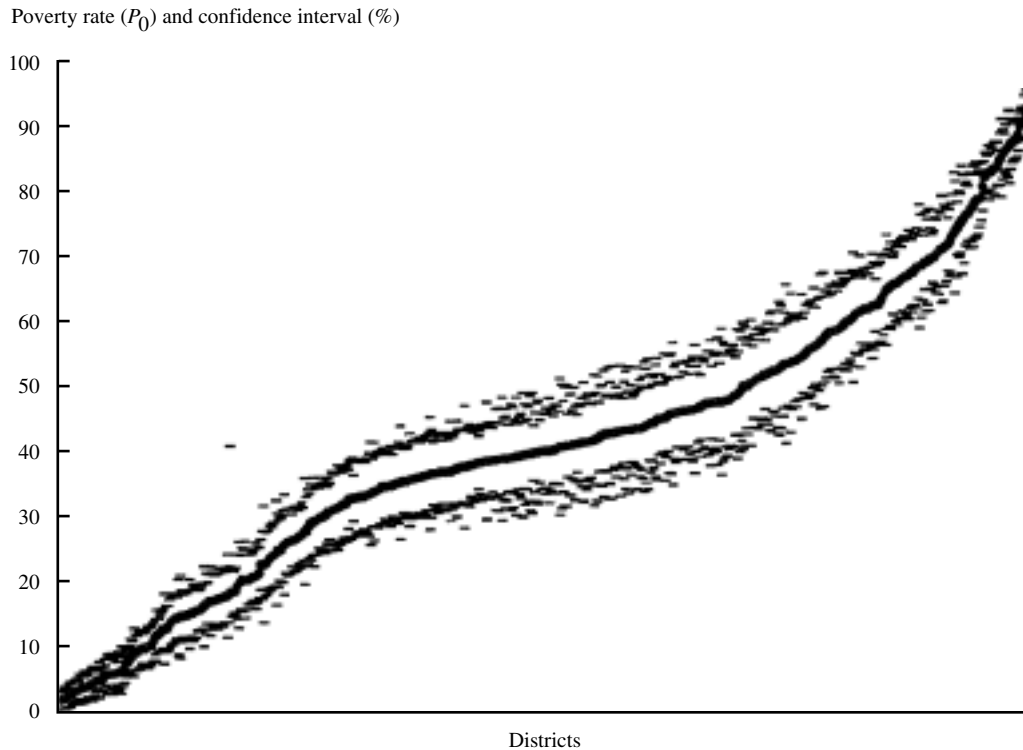
It should be noted that the sharp line between the low poverty rates in the Southeast (as defined in 1998) and the higher rates in the Central Highlands and South Central Coast are partly an artificial result of the use of regional dummy variables in Stage 1.

The 95 percent confidence intervals for the district-level poverty rates are shown in Figure 3.4. As in Figure 3.2, the center line represents the estimates of the poverty rate, and the small horizontal lines above and below the center line are the upper and lower 95 percent confidence limits.

The district-level confidence intervals range from  $\pm 1.3$  to  $\pm 22$  percentage points, with an average value of  $\pm 5.8$  percentage points. Half of the districts have confidence intervals between  $\pm 4.4$  and  $\pm 6.9$  percentage points (this is the interquartile range). As noted before, the confidence intervals are smaller (and the poverty rate estimates more accurate) when the poverty rate is close to zero or close to 100 percent. When the poverty rate is in the middle range (40–50 percent), the confidence intervals tends to be  $\pm 5$  to  $\pm 10$  percentage points.

In general, the confidence intervals for district poverty rates are somewhat higher than those of provincial poverty rates, for which the average was  $\pm 5.2$  percent. This is because there are fewer households in the districts than in the provinces. The least reliable district estimate is for Bach Long Vi district in Hai Phong province: the poverty rate is estimated as 19 percent  $\pm 22$  percent. The reason this poverty estimate is very unreliable is that it is based on just 18 households on this island district. This is an exception, however. The second highest confidence interval is  $\pm 12$  percentage points. Furthermore, only 6 of the 614 districts have fewer than 1,000 households in our sample of the Census data, and 90 percent of them have more than 2,500 households.

As noted above, urban poverty rates are generally lower than rural poverty rates. Most district-level urban poverty rates are clustered in the range of 10 to 30 percent,

**Figure 3.4 District poverty rates and confidence intervals**

and district-level rural poverty rates range from less than 10 percent to over 90 percent, with the bulk of the districts falling in the range of 30–60 percent. It is also interesting to note that there is a positive correlation ( $R^2 = 0.42$ ) between the urban poverty rate in a district and the rural poverty rate in the same district, as shown in Figure 3.5.

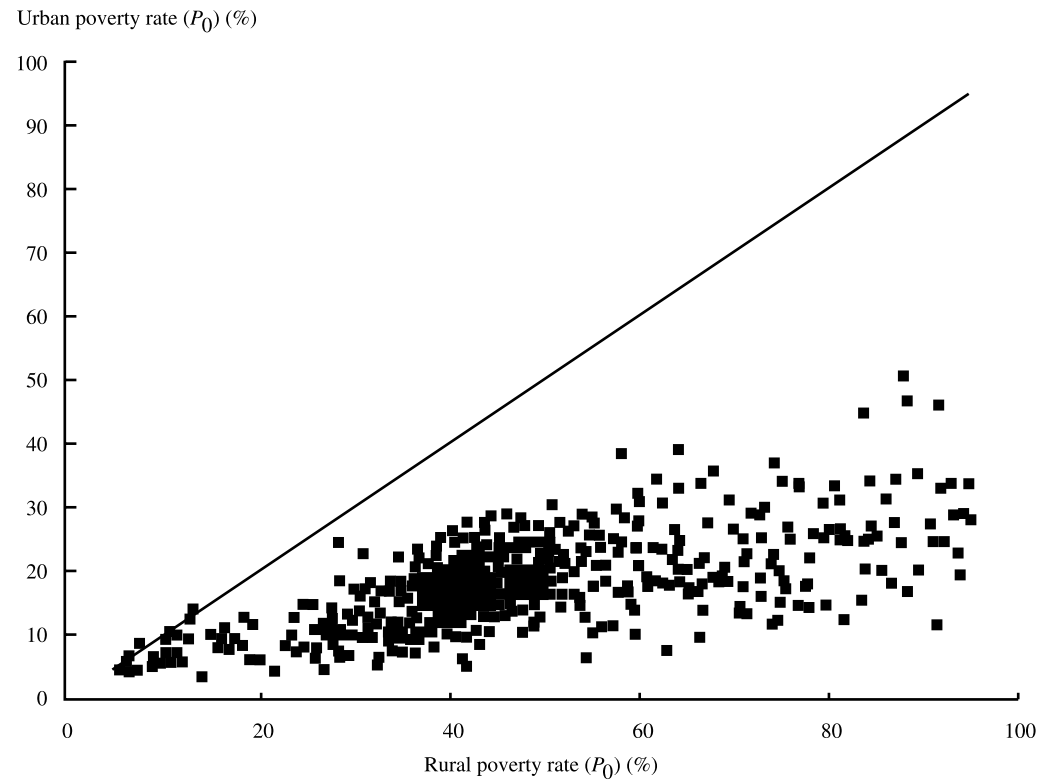
### Commune Poverty Rates ( $P_0$ )

The poverty-mapping method also generates estimates of the incidence of poverty ( $P_0$ ) at the commune level. It is important to use these results with caution because the small number of households in some communes means that the poverty estimates are not reliable for some communes. The reliability of the commune poverty estimates is discussed further below.

The spatial patterns in commune poverty rates are shown in Figure 3.6 (see color insert). This map provides considerably more detail than the district poverty map (see Fig. 3.3 in color insert). For example, in the

Northern Upland region, the commune poverty map reveals a number of green “dots” in the orange and red areas. These are urban areas with relatively low incidence of poverty surrounded by rural areas with much higher poverty rates. In addition, the path of the Red River can be seen as a yellow and orange line entering Vietnam from the northwest and heading toward Hanoi. Communes near the Red River benefit from flat land, irrigation water, and transportation provided by the river, all of which reduce poverty rates.

In the Northeast, the influence of the road network is visible in some places. For example, there is a yellow-orange line extending southeast from the Chinese border through Cao Bang and Lang Son. This corresponds to the path of the highway (Routes 4A and 4B) that goes from the Chinese border through the two provincial capitals to the coast. This may reflect the impact of market access on poverty rates, or it may be that the roads are built in less mountainous

**Figure 3.5 Urban poverty and rural poverty by district**

areas with greater agricultural potential (compare Figures 3.6 and 3.7 in color insert).

The city of Hanoi shows up clearly as a dark green dot surrounded by light green and yellow. This reflects the fact that the communes of Hanoi have poverty rates below 10 percent, compared to 20–40 percent in most of the surrounding communes of the Red River Delta.

In the North Central Coast, the commune poverty map illustrates even more clearly that the coastal communes have lower poverty rates, whereas the interior regions along the Lao border have much higher poverty rates. One exception to this pattern is the area west of the city of Vinh. The lower poverty rate in this region may reflect the impact of the highway (Route 8) that runs from Vinh west to the Lao border and serves as an important channel for trade between the two countries. Another highway (Route 7) running northwest from Vinh to the Lao border in Nghe An is barely vis-

ible in Ky Son district (see Figs. 3.6 and 3.7 in color insert).

The South Central Coast has some of the poorest coastal areas in Vietnam. In particular, the coast of Binh Thuan and Ninh Thuan provinces are shaded orange on the map, indicating poverty rates over 60 percent. The high incidence of poverty in this region is probably because this is one of the most arid parts of Vietnam, and the sandy soils make it difficult to practice intensive agriculture. A relatively large proportion of the population is involved in fishing in this area (see Fig. 3.6 in color insert).

In the Central Highlands, the commune map shows more clearly that the incidence of poverty is highest in the northern part of the Central Highlands, particularly in the provinces of Gia Lai and Kon Tum. The least poor communes in these two provinces (shaded in green and yellow) are along the north-south highway connecting Kon Tum town and Pleiku (Route 14) and along the

east-west road connecting Pleiku and the coast (Route 19). Dak Lak is less poor than Gia Lai and Kon Tum, most of the communes being shaded yellow or green indicating poverty rates below 60 percent. The town of Buon Ma Thuot is visible as a green dot in the center of the province. Dak Lak is the main coffee-growing area in Vietnam and has benefited from the rapid growth in coffee production during the 1990s (see Figs. 3.6 and 3.7 in color insert).

The commune-level poverty map also shows red areas on three sides of Lam Dong province in the Southeast region. These correspond to mountainous areas with difficult access to roads and markets. The green and yellow path in the center of the province follows the highway (Route 20) northeast from Ho Chi Minh City to Dalat, the dark green area. After Dalat, the highway turns southeast to the coast, barely visible on the map from the yellow and orange communes along the route. The Southeast region also has a large green area indicating relatively low levels of poverty (under 20 percent) in most of the provinces near Ho Chi Minh City (see Figs. 3.6 and 3.7 in color insert). As noted earlier, the sharp line between low poverty rates in the Southeast and higher rates elsewhere is partly a result of the use of regional dummy variables. Although the average poverty rate for the Southeast closely matches the VLSS poverty rate for the region, the poverty rates for outlying communes in the Southeast may be underestimated.

The Mekong Delta, in contrast, is predominantly light green and yellow, suggesting poverty rates in the range of 30–60 percent. A few communes in this region have poverty rates above 60 percent, represented by the light orange communes. These communes are found in Tra Vinh and Soc Trang, near the mouth of the Tien Giang and Hau

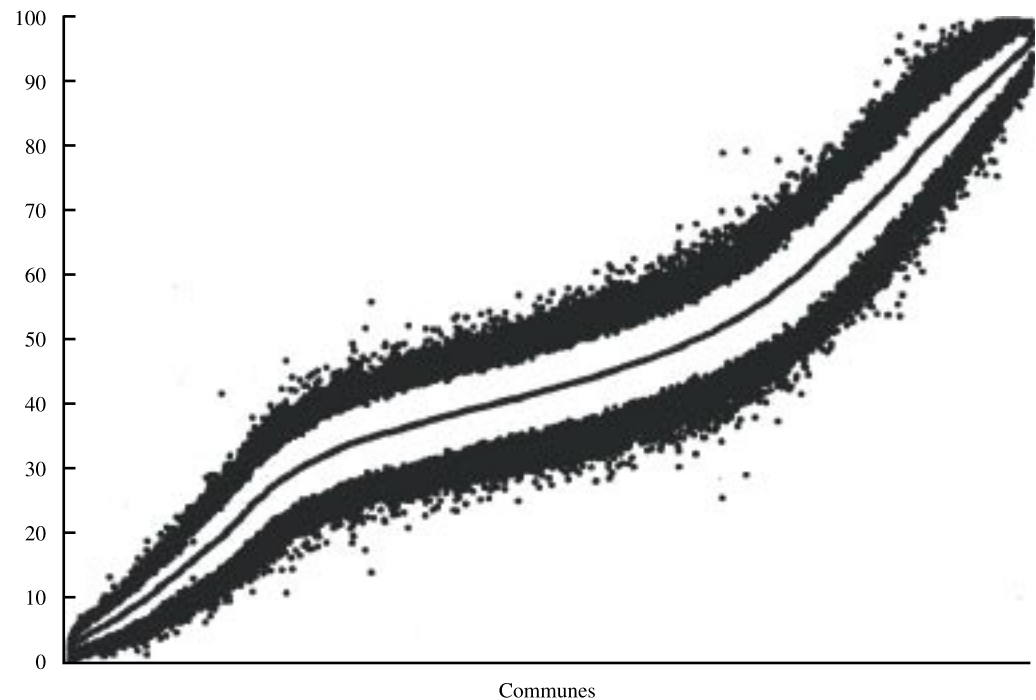
Giang branches of the Mekong River, and in An Giang and Kien Giang near the Cambodian border.

As mentioned earlier, the commune-level estimates of poverty must be interpreted with caution. Many of the communes have a relatively small number of households, leading to relatively high margins of error in the poverty estimates (see Fig. 3.8). The 95 percent confidence intervals for commune-level poverty estimates range from less than  $\pm 1$  percentage point to  $\pm 27$  percentage points, the average being  $\pm 8.1$  percentage points. Half of the communes have confidence intervals between  $\pm 6.6$  and  $\pm 10$  percentage points. This means that one-quarter of the communes have confidence intervals greater than  $\pm 10$  percentage points. By comparison, none of the province-level confidence intervals was greater than  $\pm 10$  percentage points, and only 15 of the 614 district-level confidence intervals were this large. Clearly, the commune estimates of poverty must be used very cautiously, taking into account the size of the confidence intervals. For some communes, they should not be used at all.

### Poverty Density

The three maps presented in Figures 3.1, 3.3, and 3.6 (see color insert) show the incidence of poverty, defined as the percentage of the population living below the poverty line. Another way to look at the spatial distribution of poverty is to examine the poverty density, defined as the number of poor people living in a given area. By multiplying the commune-level poverty rates by the population in each commune, we estimate the number of poor people living in that commune, a number that is represented by the number of dots in that commune.<sup>14</sup> Figure 3.9 shows the poverty density in Vietnam, where each dot represents 500 poor people.

<sup>14</sup>We do not have information on the geographic distribution of households within each commune, so the dots are distributed randomly within each commune.

**Figure 3.8 Commune poverty rates and confidence intervals**Poverty rate ( $P_0$ ) and confidence interval (%)

It is somewhat surprising to find that the number of poor people per square kilometer is greatest in the Red River Delta, in the Mekong River Delta, and along the coastal plains. The poverty density is lowest in the areas where the incidence of poverty is the highest. This is because the areas with the highest poverty rate tend to be remote and sparsely populated areas, and the lower population density more than offsets the higher percentage of the population that is poor.

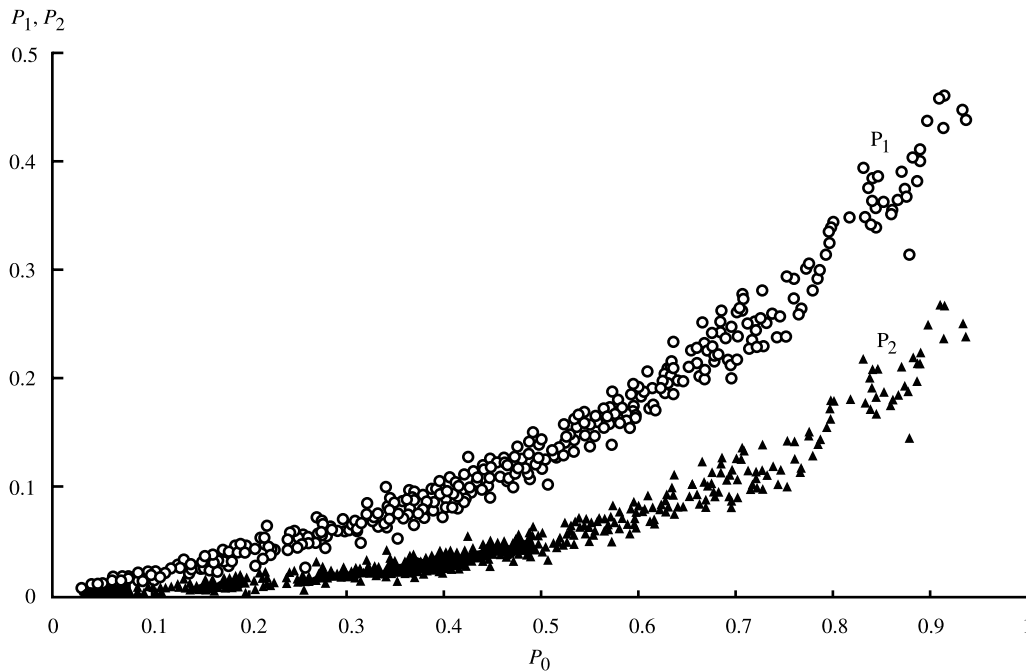
An important implication of Figure 3.9 (see color insert) is that if all poverty alleviation efforts are concentrated in the areas where the poverty rate is the highest, including the Northeast, the Northwest, the Central Highlands, and the interior of the central coast, most of the poor will be excluded from the benefits of these programs. The implications of this map are discussed in Chapter 6.

### Spatial Patterns in Other Measures of Poverty

The previous section explores the spatial patterns in the incidence of poverty, also called  $P_0$ . As described in Chapter 2, there are other measures of poverty that have useful properties. The depth of poverty ( $P_1$ ), also called the poverty gap, takes into account not just how many people are poor but how poor they are, on average. In fact, the depth of poverty is equal to the proportion of the population that is poor multiplied by the percentage gap between the poverty line and the average per capita expenditure of the poor. The severity of poverty ( $P_2$ ), also called the poverty gap squared, takes into account not just how poor the poor are, on average, but the distribution of income among them (see Chapter 2 for more information).

At the national level, the estimated value of  $P_1$  is 0.100, implying that the average poor person has a level of per capita

**Figure 3.11** Depth of poverty ( $P_1$ ) and severity of poverty ( $P_2$ ) as a function of the incidence of poverty ( $P_0$ ) in each district



expenditure that is 28 percent below the poverty line.<sup>15</sup> The estimated value of  $P_2$  at the national level is 0.040.

Figure 3.10 (see color insert) shows the district-level maps of the depth of poverty ( $P_1$ ) and the severity of poverty ( $P_2$ ), presented side-by-side to make comparison easier. It is obvious that the spatial patterns in  $P_1$  and  $P_2$  are quite similar to each other and similar to the spatial pattern of  $P_0$  (see Fig. 3.3 in color insert). In all three maps, poverty is greatest in the Northwest, Northeast, the interior of the North Central Coast, and in the northern part of the Central Highlands. Poverty is intermediate in the Red River Delta and the Mekong River Delta, and it is lowest in the large urban areas such

as Hanoi and Ho Chi Minh City as well as in the Southeast region near Ho Chi Minh City.

Figure 3.11 plots the depth of poverty ( $P_1$ ) and the severity of poverty ( $P_2$ ) on the vertical axis with the incidence of poverty ( $P_0$ ) on the horizontal axis, with each point representing one district. As the incidence of poverty rises, the depth and severity of poverty rise as well. The correlation between the poverty measures is quite strong.<sup>16</sup> The fact that the  $P_1$  line curves upward as  $P_0$  increases implies that, as the poverty rate rises, the percentage gap between the poverty line and the per capita expenditure of the average poor households increases as well.<sup>17</sup>

<sup>15</sup> $P_1 = P_0 \cdot G$  where  $G$  is the gap between the poverty line and the average per capita expenditure of poor people, expressed as a proportion of the poverty line.

<sup>16</sup>A quadratic trend line based on  $P_0$  has an  $R^2$  of 0.98 in the case of  $P_1$  and 0.96 in the case of  $P_2$ .

<sup>17</sup>If the average poverty gap remained constant,  $P_1$  would have a positive and linear relationship with  $P_0$ . The fact that it curves upward implies that the average poverty gap must also be increasing as we move from less poor to poorer districts.

## Spatial Patterns in Inequality

As discussed in Chapter 2, the small-area estimation method is most commonly used to estimate the incidence of poverty (poverty mapping), but it can also be used to generate inequality estimates. Whereas poverty measures focus on those below the poverty line, inequality measures look at the distribution of the entire population, poor and nonpoor. In this analysis, we focus on three commonly used measures of inequality: the Gini coefficient, the Theil L index of inequality, and the Theil T index of inequality. The two Theil indexes are also part of a class of generalized entropy measures, sometimes labeled GE(0) and GE(1).

### Gini Coefficient

The Gini coefficient is a measure of inequality that varies between 0 (when everyone has the same expenditure or incomes) and 1 (when one person has everything!). Thus, a higher Gini coefficient implies more inequality. For most developing countries, Gini coefficients range between 0.3 and 0.6. According to our analysis, the national Gini coefficient is 0.323, indicating a relatively low degree of inequality in per capita expenditure.

Like other measures of inequality, the Gini coefficient tends to be smaller for smaller areas, such as provinces or districts, than for the nation as a whole. This is because households in a small area are likely to be more similar to each other than to households across the entire country. Figure 3.12 (see color insert) shows the level of inequality in per capita expenditure as measured by the Gini coefficient at the district level. The areas with the least inequality (shaded white) include the Red River Delta, some lowland areas of the Northeast, coastal districts in the North Central Coast region, some districts in the Mekong Delta, and scattered coastal districts in the South Central Coast. The greatest level of expenditure inequality is found in the large urban areas, particularly Hanoi and Ho Chi Minh City,

and in the upland areas, including the Northeast, Northwest, and Central Highlands.

It is not surprising that large urban areas have a high levels of inequality because they have some of the richest households in the country as well as recent immigrants and others whose income is barely higher than in rural areas. Nor is it surprising that inequality is low in the Red River Delta and coastal districts. These areas are characterized by intensive irrigated agriculture and a large percentage of the population depending on agriculture. The agricultural potential of the irrigated farm land is relatively uniform, and the allocation of cooperative land among households was carried out with the objective of maintaining equality among households.

The Mekong River Delta is also characterized by intensive irrigated agriculture and a large percentage of the population depending on agriculture, but there is greater variation in farm size as well as the presence of some landless households that depend on selling agricultural labor.

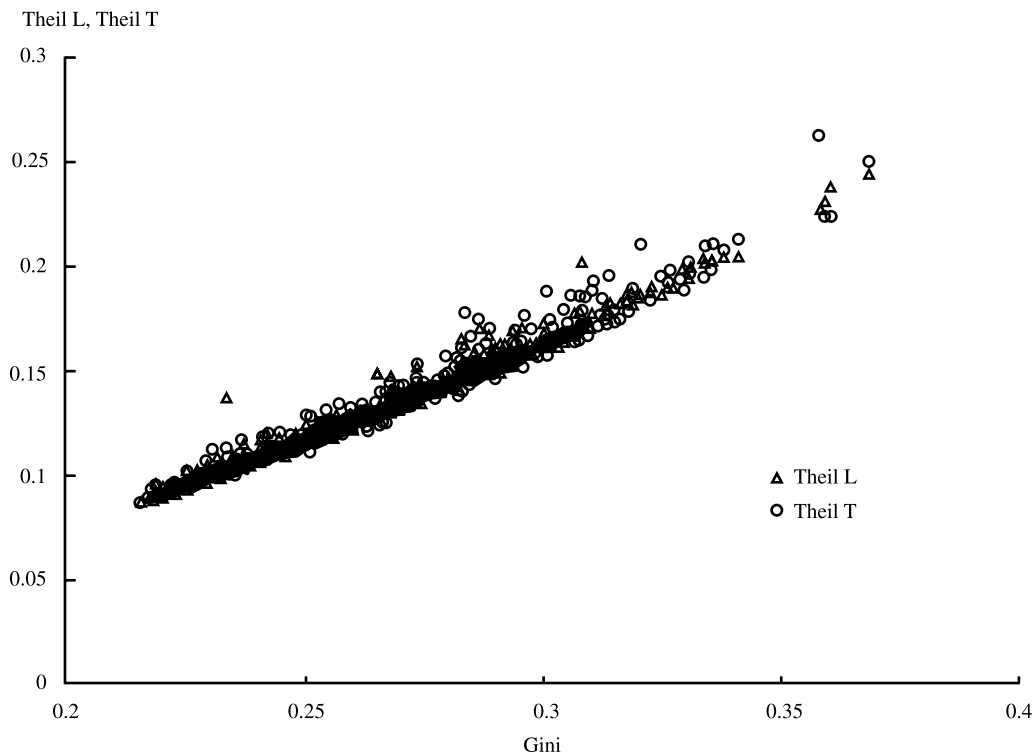
Perhaps the most surprising finding is the high level of inequality in some parts of the Northeast, Northwest, and the Central Highlands. One possible explanation is that these areas combine very poor subsistence farmers, many of whom are ethnic minorities, and some richer households who earn income from commerce, commercial agriculture (including livestock production), or salaried employment, including government employment. The case of Dak Lak, however, does not support this explanation. We would expect the contrast between relatively rich coffee farmers and poor subsistence farmers to yield a particularly high level of inequality, but Dak Lak is one of the few upland provinces where inequality is moderate.

### Theil L and Theil T Indexes of Inequality

The Theil L index varies between 0 (absolute equality) and infinity ( $\infty$ , absolute inequality), although it is unusual for it to exceed 1. Like the Gini coefficient, a higher



**Figure 3.14** Theil indexes of inequality as a function of the Gini coefficient for each district



Theil index implies a more unequal distribution of expenditures (or incomes). However, the Theil L gives more weight to the bottom of the distribution, thus giving greater weight to the distribution of expenditure among the poor than the Theil T index or the Gini coefficient.

The Theil T index varies between 0 and  $\log(N)$ , where  $N$  is the population. Unlike the Theil L index, the Theil T index gives equal weight to all parts of the distribution. The equations used to calculate the two Theil indices are given in Chapter 2.

The district-level maps of inequality as measured by the Theil L and Theil T indexes are shown in Figure 3.13 (see color insert). Despite their different underpinnings, the maps of inequality using the two Theil indexes give similar results to the map using the Gini coefficient. In all cases, inequality is lowest in the Red River Delta and some coastal districts in the North Central Coast

region, intermediate in the Mekong River Delta, and greatest in urban areas, the Northern Uplands, and the Central Highlands.

Figure 3.14 shows the relationship among the three measures of inequality, where each dot represents one district. This graph indicates that there is a linear relationship between the Gini coefficient on the one hand and the two Theil indexes on the other and that the correlation is quite close. For the relationship between the Gini coefficient and the Theil L index,  $R^2 = 0.98$ , and in the relationship between the Gini coefficient and the Theil T index,  $R^2 = 0.97$ . This helps to explain why the three inequality maps are quite similar.

### Decomposing Inequality

Is inequality mainly caused by differences across provinces or differences across households within each province? Unlike the Gini coefficient, the Theil L and T indexes

**Table 3.5 Decomposition of inequality into between- and within-province components**

Inequality measure	Variable	Total inequality at national level	Between-province component	Within-province component
Theil L index	Value of index	0.193	0.046	0.147
	Share of total	100%	24%	76%
Theil T index	Value of index	0.204	0.050	0.153
	Share of total	100%	25%	75%

**Table 3.6 Decomposition of inequality into between- and within-district components**

Inequality measure	Variable	Total inequality at national level	Between-district component	Within-district component
Theil L index	Value of index	0.193	0.067	0.127
	Share of total	100%	34%	66%
Theil T index	Value of index	0.204	0.073	0.131
	Share of total	100%	36%	64%

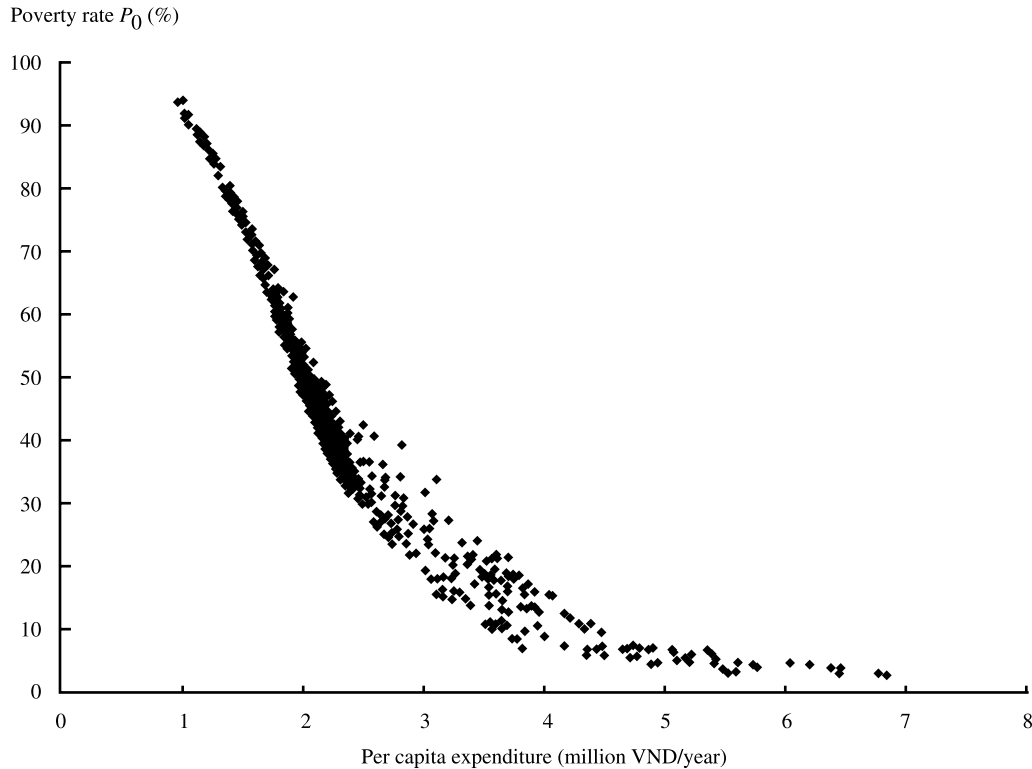
of inequality can be exactly decomposed into “subgroups.” For example, the Theil index for all of Vietnam is equal to the weighted average of the provincial indexes (the “within-province” component) plus the Theil index of the inequality in provincial average expenditures (the “between-province” component). The “between” component refers to what inequality would be if everyone inside a province had the same expenditure as the provincial mean, and the “within” component takes into account inequality within provinces but excludes inequality of provincial means.

Table 3.5 decomposes the Theil L and Theil T measures using provinces as the subgroup. The between-province component of inequality accounts for about one-quarter of the inequality at the national level. The other three-quarters results from inequality within each province. The magnitude of these decomposition results are similar to those described by Kanbur (2002) for other developing countries, where it is

usual for the between component to account for around 15 percent of total national inequality.<sup>18</sup>

Table 3.6 decomposes the Theil L and T measures using districts as the subgroup. The between-district component of inequality is about one-third (34 percent for the Theil L index and 36 percent for the Theil T index). The other two-thirds of national inequality is associated with inequality within each district. We expect the between component of inequality to increase with greater geographic disaggregation. It is perhaps more surprising how much inequality remains after disaggregating down to the level of the 614 districts. This suggests that district-level targeting in antipoverty programs may not be that effective, though more detailed studies of leakage and under-coverage rates would be needed to confirm this conclusion. These results also contradict the widespread view in Vietnam that inequality between provinces is a major contributor to overall inequality.

<sup>18</sup>Obviously the relative magnitude of the between and within components will depend on how many subgroups (provinces, districts, or other administrative units) are involved. The greater the number of subgroups, the larger is the between component.

**Figure 3.15 Poverty rate ( $P_0$ ) as a function of per capita expenditure**

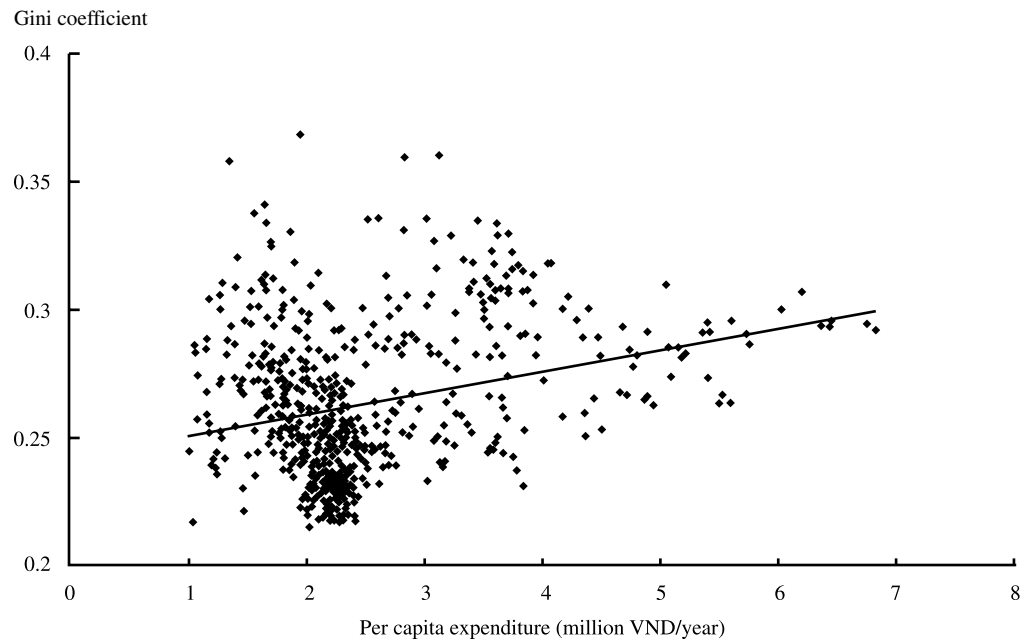
### Relationships among Income, Poverty, and Inequality

Previous sections of this chapter examined the spatial patterns in poverty and inequality. In this section, we examine the relationship among poverty, inequality, the degree of urbanization, and average per capita expenditure at the district level. In order to reduce the number of variables, and because of the close correlation among poverty measures, we use  $P_0$  to represent poverty. Similarly, because all three inequality measures are closely correlated, we use the Gini coefficient to represent inequality.

In Figure 3.15, we plot the poverty rate ( $P_0$ ) as a function of the district average per capita expenditure, where each dot represents a district. We expect that as per capita expenditure rises, the poverty rate will fall. Nonetheless, it is surprising how closely the poverty rate depends on the average per capita expenditure of the district. Particularly

among the poorer districts, the relationship between the two is very close. A quadratic trend line explains 96 percent of the variation in poverty. This suggests that the incidence of poverty in a district is largely a function of the average level of per capita expenditure in the district and that the degree of inequality within a district plays a minor role in determining the poverty rate.

Figure 3.16 shows the relationship between the Gini coefficient and the average per capita expenditure of the district. It is widely believed in Vietnam and other countries that as incomes rise, the gap between the poor and rich widens. The data presented here confirm that view to some degree. The linear trend line shown on the graph indicates an increase in the Gini coefficient from 0.25 to 0.30 as per capita expenditure rises from 1 million VND/year to 7 million VND/year. This may be part of the pattern found in international data in which, at low levels of income, higher income is associ-

**Figure 3.16** Gini coefficient of inequality as a function of per capita expenditure

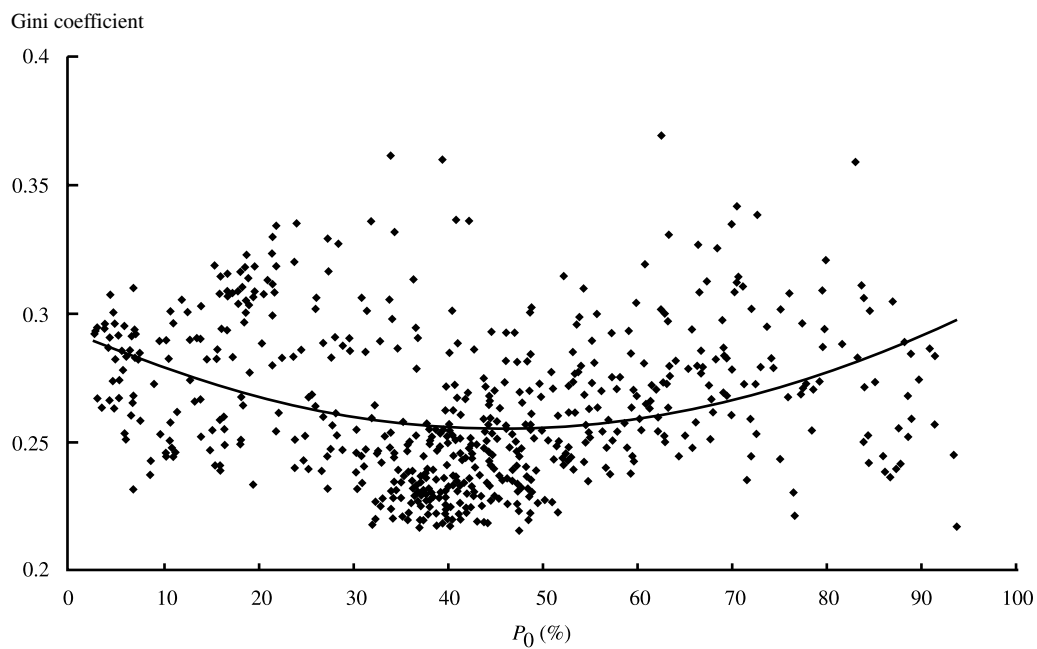
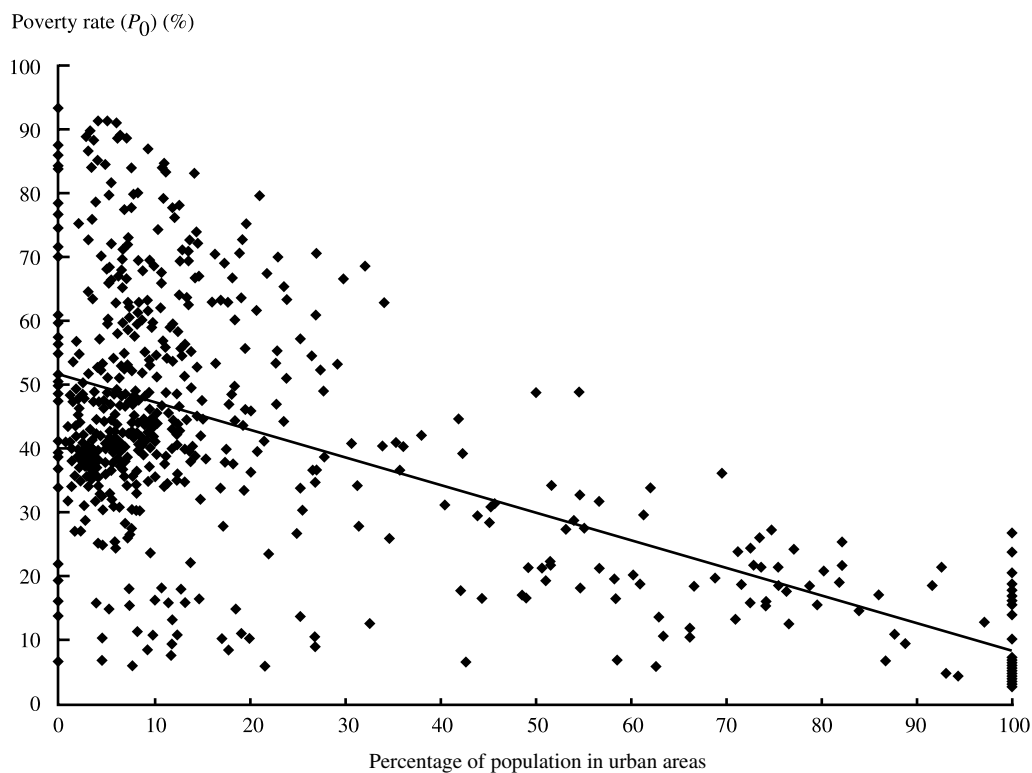
ated with higher inequality, but at some point further increases in income tend to reduce inequality. This inverted-U pattern is called the Kuznets curve. Because Vietnam is a low-income country, the Kuznets curve would predict a positive relationship between income and inequality over time and across districts.

But the relationship between inequality and per capita expenditure in Figure 3.16 is not a simple positive relationship. Many low-income districts also have a high level of inequality. In fact, the districts with the highest levels of inequality tend to be the relatively poor districts with per capita expenditure below 4 million VND/year. Furthermore, low-income districts have a wider range of levels of inequality, whereas high-income districts seem to converge toward a Gini coefficient of around 0.3.

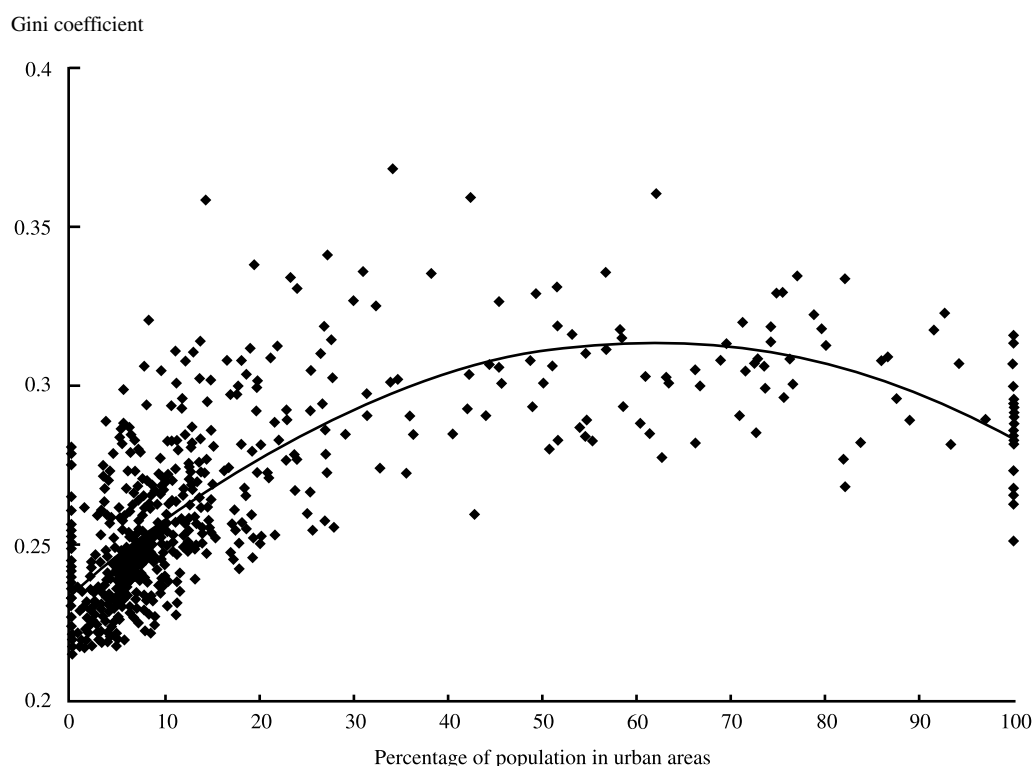
The relationship between poverty ( $P_0$ ) and inequality (the Gini coefficient) is shown in Figure 3.17. There appears to be a weak U-shaped pattern in which the highest level of inequality is found in the poorest districts and in the least poor districts. This may be related to the pattern noted in Figures 3.12

and 3.13 (see color insert), in which the areas with the highest inequality were the rural upland areas (with high poverty rates) and the large urban centers (with low poverty rates). Although the trend line does support the idea of a quadratic (curved) relationship, the relationship is fairly weak ( $R^2 = 0.12$ ).

Rural poverty rates exceed urban poverty rates in almost every country where it has been studied. Indeed, this pattern has been confirmed for Vietnam by various surveys (for example, GSO 2000). Using small-area estimation methods, however, we can examine the poverty rates for many urban and rural districts to provide a more detailed picture of the relationship between the degree of urbanization and poverty. Figure 3.18 shows the relationship across districts between the proportion of the population living in urban areas and the poverty rate ( $P_0$ ), along with a linear trend line. The graph indicates clearly that there is a clear negative relationship: most districts that are largely rural have poverty rates in the range of 30–60 percent, whereas most districts that are mainly urban have poverty rates that

**Figure 3.17** Gini coefficient of inequality as a function of the poverty rate ( $P_0$ )**Figure 3.18** Poverty rate ( $P_0$ ) as a function of the share of the population in urban areas

**Figure 3.19** Gini coefficient of inequality as a function of the share of the population in urban areas



are less than 30 percent. At the same time, it is interesting to note the wide range of poverty rates among rural districts. Several dozen districts have a majority rural population and poverty rates in the same range as urban districts. This suggests that under some circumstances, poverty can be reduced significantly within rural areas. Based on the maps presented earlier, it is clear that many of these “rich” rural districts are in the Southeast region, benefiting from the access to labor and commodity markets in Ho Chi Minh City.

The relationship between the degree of urbanization and inequality is quite different. As shown in Figure 3.19, inequality (as measured by the Gini coefficient) is quite low for districts that are almost entirely rural, and it is almost as low for districts that are almost entirely urban. The districts with the highest level of inequality are those that combine urban and rural populations,

with the urban share of the population being in the range of 20 to 80 percent. These results confirm the common view that urban areas have more inequality than rural areas, but it suggests that the pattern is more complicated in that districts with both rural and urban populations have the highest inequality. The quadratic trend line shows that inequality is at its highest when the urban share is about 60 percent.

### **Relationship with MOLISA Poverty Estimates**

In this section, we compare the estimates of the incidence of poverty ( $P_0$ ) derived from our application of the small-area estimation method to the estimated incidence of poverty produced by the Ministry of Labor, Invalids, and Social Affairs (MOLISA). As described earlier, there are a number of differences in the definition of poverty and the

data collection methods. Some of these differences are summarized below:

- Our definition of poverty uses as the welfare indicator the value of per capita consumption expenditure, including the value of subsistence food production and the imputed rental value of owner-occupied housing. In contrast, MOLISA uses per capita income as its welfare indicator.
- To adjust for regional differences in the cost of living, we use a set of regional and monthly price indexes calculated by the GSO for the 1997–98 VLSS analysis. These price indexes are based on the cost of a basic consumption basket in the urban and rural areas of each region. In contrast, MOLISA adjusts for the local cost of living by expressing per capita income in terms of the number of bags of rice it will buy at local prices.<sup>19</sup>
- Our poverty line is equal to the “overall poverty line,” defined as VND 1,789 million per person per year in real consumption expenditure. MOLISA defines the poverty line in terms of the number of bags of rice, although the number varies somewhat from one province to another.
- We define the poverty rate in terms of the percentage of people living in households whose per capita expenditure is below the poverty line. MOLISA defines the poverty rate in terms of the percentage of households below the poverty line.
- Our poverty estimate for each district is based on the characteristics of households in that district in the 1999 Population and Housing Census, given the relationship between per capita expenditure and those household characteristics in the 1997–98 Vietnam Living Standards Survey. The MOLISA estimates are based on assessments of

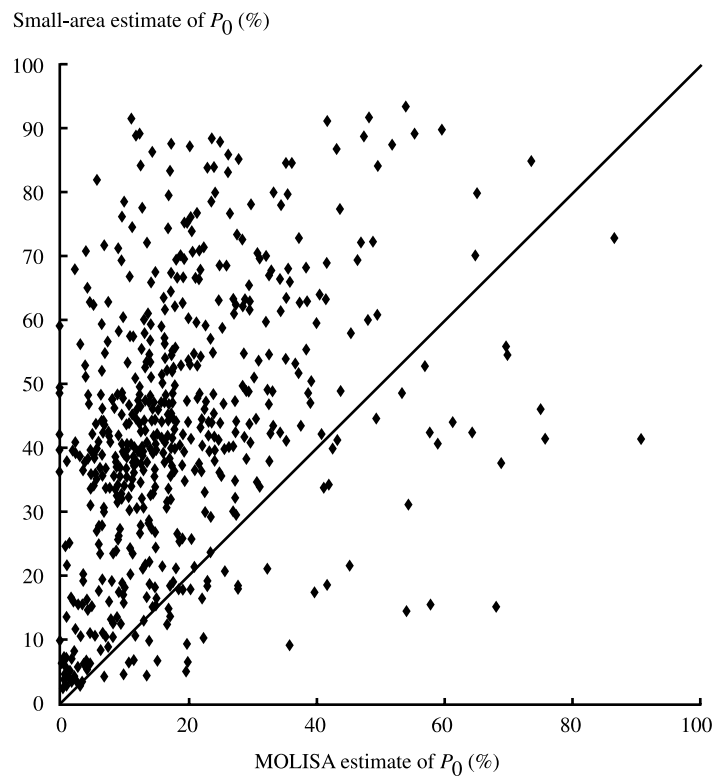
MOLISA field staff in each commune, applying national and provincial guidelines to identify poor households (for a description of the field work, see Conway 2001).

Do these methodological differences result in different estimates of the incidence of poverty ( $P_0$ ) at the district level? As shown in Figure 3.20, the MOLISA poverty estimates are generally lower than those generated by the small-area estimation method used in this report. The median value of the MOLISA poverty rates is 15 percent, compared to 41 percent for our poverty estimates. This difference is not particularly surprising because the two estimates are based on quite different poverty lines. Given the wide range of views about how to construct a poverty line, there is little to be gained from debates over the “true” poverty rate. It is more important to examine whether the spatial patterns in poverty are consistent between the two methods.

What is surprising is that there is very little correlation between the district-level poverty rate estimates produced by MOLISA and the poverty rates estimated by this study (the  $R^2$  of a linear trendline is just 0.17). To illustrate the disagreement in the estimates, we consider two districts in which the contrast between the two methods is the greatest. In the upper left corner of Figure 3.20 is a dot representing Bat Xat district, located in the northwest corner of Lao Cai. According to our estimates, Bat Xat district has a poverty rate of almost 82 percent. By contrast, the MOLISA poverty estimate for the district is less than 6 percent. Given that Bat Xat is in a remote portion of one of the poorest provinces in Vietnam, we would expect the poverty rate to be relatively high. In the lower right corner of Figure 3.20 is the urban district of Nha Trang, on the central coast of Khanh Hoa province. The MOLISA poverty estimate for Nha Trang is 68 per-

<sup>19</sup>This is equivalent to a price index that includes just one commodity, rice, in the consumption basket.

**Figure 3.20 Comparison of poverty rates ( $P_0$ ) from MOLISA and from small-area estimation methods**



cent, whereas the estimate produced in this report is just 15 percent. Because Nha Trang is a popular beach resort town, benefiting from both local and international tourism, a relatively low poverty rate would be expected.

Clearly, the choice of poverty estimates can make a large difference in terms of the

targeting of poverty alleviation programs. Further research is needed to resolve the discrepancies between these two poverty estimates. One approach would be to select districts where the two estimates vary widely (such as the two cited above) and collect primary or secondary data to determine which estimate conforms more closely to reality.



## CHAPTER 4

---

### Geographic Determinants of Poverty

**T**he poverty maps that have been presented in Chapter 3 show considerable geographic variation among provinces, districts, and communes in Vietnam. In particular, the incidence of poverty is highest in the upland areas of Vietnam bordering China and Laos and lowest in large urban centers and in the Red River and Mekong deltas. This chapter uses the district-level poverty estimates from Chapter 3 to investigate the extent to which geographic variables may have an effect on the incidence of poverty in a district.<sup>20</sup>

#### Geographic Factors

Table 4.1 lists a number of geographic variables that may help to explain the spatial patterns in poverty in Vietnam. The variables are divided into two categories. Exogenous variables are those that are unlikely to be affected by the level of economic activity or poverty. For example, agro-climatic variables such as rainfall or topography may influence poverty, but they are unlikely to be influenced by poverty. In contrast, the endogenous variables may both influence poverty and be influenced by it (at least in the long run). For example, areas with low poverty rates may attract immigrants, increasing the population of the area. Similarly, investments in markets and transport infrastructure is determined at least in part by the level of economic activity, so that a low poverty rate may influence the density of markets and roads in the long run.

The right-side column of Table 4.1 shows the expected relationship between each variable and poverty. The double-sided arrows indicate cases in which poverty and the variable each have some effect on the other. In order to carry out a regression analysis, the agro-climatic factors in Table 4.1 must be expressed as specific variables.

Using a Chow test, we determined that the coefficients explaining urban and rural poverty were significantly different from each other, indicating that separate urban and rural models would be preferable. We would expect soil type, land cover, rainfall, and sunshine to matter much more in rural areas, where approximately two-thirds of households list agriculture as their main economic activity. In contrast, indicators of nonagricultural activities, such as the number of markets per district and distance to major cities, are expected to be more important in urban areas.

---

<sup>20</sup>We decided not to carry out the analysis of geographic determinants of poverty at the commune level for two reasons. First, the commune-level poverty estimates have large standard errors, indicating a large “noise” component in these estimates. Second, some of the geographic variables are less accurate at the commune level. For example, climate variables are interpolated from a relatively small number of weather stations. Interpolation at the district level is probably more reliable than interpolation at the commune level.

**Table 4.1 Agro-climatic and socioeconomic factors that may affect poverty rate**

Variables	Expected relationship to poverty
Exogenous variables	
Elevation	Higher elevation → higher poverty
Slope/roughness	Steeper slopes → higher poverty
Soil type	Sandy and poor soils → higher poverty
Type of land cover	Not known
Hours of sunshine	Less sunshine → higher poverty
Rainfall per year	Low rainfall → higher poverty
Distance to towns and cities	Greater distance → higher poverty
Endogenous variables	
Population	Not known
Number and density of markets	Lower density of markets ↔ higher poverty
Length and density of roads	Lower density of roads ↔ higher poverty
Length and density of navigable rivers	Lower density of navigable rivers ↔ higher poverty
Transport time to towns and cities	Higher transport time ↔ higher poverty

### Estimation Issues

As discussed in Chapter 2, poverty rates in nearby districts are likely to be similar to one another, so it is important to pay attention to the structure of spatial dependence in our data. Failure to do this can result in inconsistent or biased estimates of the impact of different geographic variables, especially when ordinary least squares is used as the estimation method. The spatial econometrics literature distinguishes between two types of spatial dependence:

- Spatial error dependence, in which unobserved explanatory variables are correlated over space. An example of this would occur if, because of provincial policies and budgets, the quality of local health care were similar across all districts in a province but different across provinces. When there is spatial error dependence, ordinary least-squares regression coefficients will be unbiased but not efficient (the standard errors will be larger than they would be if all information were used).
- Spatial lag dependence, in which the dependent variable in one area is di-

rectly affected by the dependent variables in nearby areas. An example would be that the poverty rate in one area is directly affected by poverty in nearby districts. When there is spatial lag dependence, ordinary least-squares regression coefficients are biased and inconsistent.

Whenever spatial error or spatial lag dependence is indicated, special types of generalized least-squares (GLS) regression models need to be applied. In the case of spatial error dependence, the spatial error model is appropriate, whereas in the case of spatial lag dependence, the spatial lag model would be used.<sup>21</sup> In both cases, the researcher must specify the structure of spatial weights, which defines the functional form of the weights as a function of distance or contiguity.

Our estimation strategy is as follows. First, we estimate an ordinary least-squares (OLS) model with all exogenous variables included. Second, we perform tests for the two types of spatial dependence. Third, we use either the spatial error or the spatial lag

<sup>21</sup>See Anselin (1988) for a description of these models.

**Table 4.2 Diagnostic tests for spatial dependence in rural poverty**

Test	Statistic	df	P-value
Spatial error model			
Moran's <i>I</i>	25.459	1	0.000
Robust Lagrange multiplier	168.329	1	0.000
Spatial lag model			
Robust Lagrange multiplier	6.166	1	0.013

model to reestimate the model using generalized least squares. Two versions of the rural and urban models are presented: a more selective one that includes only strictly exogenous variables, and a more comprehensive model that includes some variables that may be endogenous, at least in the long run. We adopted spatial weights that are proportional to the inverse distance between the geographic centers of the districts, up to a maximum distance of 75 kilometers. A more detailed description of these methods and the tests is provided in Chapter 2 and in Appendix B.

### Global Model of Rural Poverty

Table 4.2 shows the tests that were conducted for spatial dependence when an ordinary least-squares model was estimated with the district-level rural poverty rate as the dependent variable and the exogenous variables listed above. Inverse-distance weights were used to perform this test.<sup>22</sup> Both Moran's *I* and the Lagrange multiplier test statistic reject the null hypothesis of no spatial dependence. The much larger Lagrange multiplier in the spatial error model indicates that this type of spatial dependence is more likely. We therefore proceed to estimate the spatial error model in order to analyze the determinants of rural poverty (see Chapter 2 for more information on these tests).

### Inclusive Model of Rural Poverty

Table 4.3 shows the results of regressing district-level rural poverty rates on the full set of unrestricted exogenous variables using the spatial error model. The model explains four-fifths of the variation in district-level rural poverty rates, a surprisingly high proportion. As in all the models presented in this section, the spatial correlation coefficient ( $\lambda$ ) is positive, large (close to 1.0), and statistically significant. This suggests that the error terms of nearby districts are strongly and positively correlated with each other (see equation 15 for the interpretation of  $\lambda$ ).

Of the 32 coefficients, only eight are statistically significant at the 5 percent level. It may seem surprising that none of the elevation variables are statistically significant, but this is probably because slope and soil type are included in the model. In other words, high elevations do not contribute to rural poverty directly but only to the extent that they are associated with poor soils and steep slopes. Among the land cover variables, only bare-rocky cover is statistically significant. The positive coefficient means that the rural poverty rate is higher in districts with a high proportion of bare and rocky land.

Two of the slope variables are statistically significant (share of the land with a 4–8 percent slope and share of the land with a 15–30 percent slope). Because the omitted category is flat land (less than 4 percent slope), these results indicate that districts

<sup>22</sup>We carried out the analysis with another commonly used form of distance weighting, inverse-squared distance, and the results were similar.

**Table 4.3 Inclusive model of the geographic determinants of rural poverty**

Characteristic	Coefficient	Robust standard error	z
Area between 251 and 500 m (%)	0.0005366	0.0004063	1.32
Area between 501 and 1,000 m (%)	0.0004463	0.0003416	1.31
Area between 1,001 and 1,500 m (%)	-0.0002176	0.0008015	-0.27
Area over 1,500 m (%)	0.0012043	0.0010897	1.11
Land cover (% of arable land)	0.0003025	0.0001767	1.71*
Land cover (% of bare and rocky land)	0.0016445	0.0005611	2.93***
Land cover (% of national forest)	-0.0000316	0.0002933	-0.11
Land cover (% of plantation forest)	-0.0017524	0.0010513	-1.67*
Slope (% of land with 4 to 8% slope)	0.0039081	0.0008831	4.43***
Slope (% of land with 8 to 15% slope)	0.0010364	0.0012501	0.83
Slope (% of land with 15 to 30% slope)	0.0039068	0.0011927	3.28***
Slope (% of land with over 30% slope)	0.0034696	0.001971	1.76*
Area with alluvial soils (%)	-0.000042	0.0002639	-0.16
Area with alluvial glacial soils (%)	0.0003011	0.0003052	0.99
Area with alluvial acidic soils (%)	0.0000304	0.000292	0.10
Area with acid sulfate soils (%)	0.000612	0.0003086	1.98**
Area with salty soils (%)	0.0005966	0.0002946	2.03**
Area with alluvial oxidized soils (%)	-0.0001836	0.0003344	-0.55
Area with red-brown soils (%)	-0.0007071	0.0003803	-1.86*
Area with sandy soils (%)	0.0012598	0.0004398	2.86***
Area with fluvial soils (%)	0.0006494	0.0003864	1.68*
Area with Acrisol soils (%)	0.0009033	0.000417	2.17**
Area with other soils (%)	-0.0000894	0.0003413	-0.26
Area with rocky soils (%)	-0.0009221	0.0006002	-1.54
Area covered with water (%)	0.0004779	0.0009002	0.53
Annual sunshine (days)	-0.000037	0.0000544	-0.68
Annual precipitation (mm)	-8.81e-06	0.0000312	-0.28
Distance from town with over 10,000 inhabitants (m)	$2.50 \times 10^{-6}$	$3.45 \times 10^{-7}$	7.26***
Distance from town with over 50,000 inhabitants (m)	$5.95 \times 10^{-7}$	$3.31 \times 10^{-7}$	1.80*
Distance from town with over 100,000 inhabitants (m)	$2.53 \times 10^{-7}$	$2.89 \times 10^{-7}$	0.88
Distance from town with over 250,000 inhabitants (m)	$8.20 \times 10^{-8}$	$2.99 \times 10^{-7}$	0.27
Distance from town with over 1 million inhabitants (m)	$-2.30 \times 10^{-8}$	$3.03 \times 10^{-7}$	-0.08
Constant	0.3901486	0.1326071	2.94***
$\lambda$	0.9015347	0.0421281	21.40***

Notes: Spatial error model:  $N = 569$ ; variance ratio = 0.679; squared correlation = 0.736; log-likelihood = 695.58012;  $\sigma = 0.07$ ;  $R^2 = 0.8009$ . Wald test of  $\lambda = 0$ :  $\chi^2(1) = 457.953$  (0.000). Lagrange multiplier test of  $\lambda = 0$ :  $\chi^2(1) = 378.471$  (0.000). Acceptable range for  $\lambda$ :  $-1.900 < \lambda < 1.000$ . \* indicates significance at the 10 percent level, \*\* at the 5 percent level, and \*\*\* at the 1 percent level.

with a large area of sloped land will have higher poverty rates than those with flat areas. This is not surprising in light of the difficulties of cultivating and irrigating sloped land as well as problems associated with erosion on steep land.

The coefficients for three soil types were also statistically significant at the 5 percent

level of confidence (because gray ferrous soils are the most common, we used this as the omitted category). For example, districts with a large share of the area covered with acid sulfate soils tend to be poorer than districts with a small share, other things being equal. Districts with large areas of salinated soils also tend to be poorer, as do districts

with sandy soils. Acid sulfate soils are a particularly serious problem in parts of the Mekong River Delta. Salination affects irrigated agricultural areas near the coast that do not have adequate infrastructure or fresh water flow to avoid contamination by seawater. Sandy soils are a problem in the South Central Coast, among other areas.

Somewhat surprisingly, annual precipitation and hours of sunshine do not have a statistically significant effect on the rural poverty rate, after controlling for other factors. It is understandable that in irrigated areas, rainfall is less of a constraint. Perhaps in rainfed areas, the problems associated with steep slopes and poor soils dominate variations in rainfall and sunshine.

Among the variables representing distance to urban centers, distance to a city of at least 10,000 people has a positive and statistically significant coefficient. This means that, other things being equal, a district that is far from a town of 10,000 inhabitants tends to be poorer than one that is close to such a town, after controlling for elevation, slope, soil type, and land cover. Distance to a town of 50,000 is not statistically significant at the 5 percent level, but it is significant at the 10 percent level. It is somewhat surprising that distance to larger cities is not statistically significant.

### Selective Model of Rural Poverty

We now turn our attention to a more selective model of rural poverty, using a combination of economic intuition and a stepwise backward regression procedure to eliminate or combine variables from the inclusive rural model. Because rainfall and sunshine are insignificant predictors of rural poverty in the unrestricted spatial error model, they are dropped. *F*-tests were then conducted, indicating that it makes sense to aggregate all four of the alluvial soils variables with those

of the omitted soil category gray ferrous soils (which make up some 37 percent of soils in Vietnam). In addition, because there is very little land with slopes over 15 percent or elevation over 500 meters, we aggregate these categories of slope and elevation with their nearest neighbors. Finally, we focus on the distance to towns with populations of over 10,000 and 100,000 and cities over one million to reduce multicollinearity with the other distance variables.

Restricting the rural poverty regression analysis in this way gives the model shown in Table 4.4 with just 12 explanatory variables. This model can still explain 74 percent of the variation in rural poverty, with poverty increasing with the slope of land or the presence of high percentages of acidic, salty, or other Acrisol soils. Bare and rocky land cover is also associated with higher levels of rural poverty, although the effect of elevation (which will be correlated with slope, soil, and land cover) is not significant. Finally, the distance to small and medium-sized towns (over 100,000 inhabitants) is positively associated with rural poverty, but distance to cities of over 1 million is not. This may be interpreted as providing evidence of the greater importance of closeness to small towns rather than major cities in reducing rural poverty.

Having developed this selective model of rural poverty, we then tested to see if including any of the endogenous variables would increase the explanatory power of the model. In theory, adding such variables may cause problems of simultaneity.<sup>23</sup> In practice, however, including endogenous variables such as population and road density in the rural poverty model only increase the explanatory power ( $R^2$ ) by 3 percentage points. Alternative specifications of these variables that replace distance to town and cities with the estimated travel time to reach

<sup>23</sup>If the explanatory variables in a regression equation are affected by the dependent variable, then coefficients estimated with ordinary least-squares regression will be biased and will not reflect the effect of changes in the explanatory variable on the dependent variable.

**Table 4.4 Selective model of the geographic determinants of rural poverty**

Characteristic	Coefficient	Robust standard error	z
Area between 251 and 500 m (%)	0.0023138	0.000644	3.59***
Area between 501 and 1,500 m (%)	0.0025204	0.0004661	5.41***
Area over 1,500 m (%)	0.0047158	0.0014319	3.29***
Area with acid sulfate soils (%)	0.0006972	0.0002769	2.52**
Area with salty soils (%)	0.0006522	0.0001998	3.26***
Area with acrisol soils (%)	0.0007839	0.0003169	2.47**
Land cover (% of bare and rocky land)	0.001079	0.000394	2.74***
Distance from town with over 10,000 inhabitants (m)	$2.75 \times 10^{-6}$	$3.26 \times 10^{-7}$	8.45***
Distance from town with over 100,000 inhabitants (m)	$6.09 \times 10^{-7}$	$2.65 \times 10^{-7}$	2.29**
Distance from town with over 1 million inhabitants (m)	$4.82 \times 10^{-8}$	$1.96 \times 10^{-7}$	0.25
Area between 251 and 500 m (%)	0.0003639	0.000395	0.92
Area over 500 m (%)	0.000278	0.0003297	0.84
Constant	0.3247383	0.0408911	7.94***
$\lambda$	0.8972043	0.0345779	25.95***

Notes: Spatial error model:  $N = 569$ ; variance ratio = 0.639; squared correlation = 0.723; log-likelihood = 676.05323;  $\sigma = 0.07$ ;  $R^2 = 0.7379$ . Wald test of  $\lambda = 0$ :  $\chi^2(1) = 673.264$  (0.000). Lagrange multiplier test of  $\lambda = 0$ :  $\chi^2(1) = 961.892$  (0.000). Acceptable range for  $\lambda$ :  $-1.900 < \lambda < 1.000$ . \*\* indicates significance at the 5 percent level and \*\*\* at the 1 percent level.

them and the population density of districts with the density of markets, produced similar results.

Finally, when dummy variables for Vietnam's seven regions are included in the rural poverty model, they are jointly significant and increase the model's explanatory power by another 9 percentage points. In many ways, however, such dummy variables are picking up the effect of omitted geographic variables and are best interpreted as fixed effects that show our inability to explain more than 75 to 80 percent of the variation in rural poverty rates using geographic variables.

It is not surprising that agro-climatic variables affect rural poverty. Poverty in rural areas is closely related to agricultural productivity and market access. The land cover, slope, and soil type have direct effects on agricultural productivity. Similarly, distance to towns and cities is one (admittedly crude) indicator of market access, which affects the prices farmers receive for their output as well as prices they pay for inputs. But it

is somewhat surprising that these variables explain such a high percentage of the variation in rural poverty across districts.

### Global Model of Urban Poverty

As with rural poverty, we first test for the type of spatial dependence when an unrestricted ordinary least-squares model is estimated with the district-level urban poverty rate as the independent variable and the exogenous variables listed above. Table 4.5 shows the diagnostic test of spatial dependence (inverse distance weights were used to perform this test). The results from the tests are now much less conclusive than before. For the global urban model with all possible exogenous variables included, there is weak evidence for preferring the spatial lag rather than the spatial error model. This can be interpreted as providing (some) evidence that poverty rates in one urban area are directly affected by poverty rates in nearby urban areas. However, the test also

**Table 4.5 Diagnostic tests for spatial dependence in urban poverty**

Test	Statistic	df	P-value
Spatial error			
Moran's <i>I</i>	18.674	1	0.000
Lagrange multiplier	192.773	1	0.000
Robust Lagrange multiplier	11.678	1	0.001
Spatial lag			
Lagrange multiplier	195.959	1	0.000
Robust Lagrange multiplier	14.864	1	0.000

indicated that the spatial error model may be appropriate, and preference for this model was confirmed separately for the more selective model (without and with endogenous variables) developed below.

### **Inclusive Model of Urban Poverty**

As before, we start with an inclusive model and then identify a more selective model. The inclusive model of district-level urban poverty explains just 38 percent of the variation in urban poverty levels (see Table 4.6). This indicates that urban poverty is much harder to explain with geographic variables alone than rural poverty.

Just 4 of the 32 explanatory variables have coefficients that are statistically significant at the 5 percent level. The percentage of district land that is arable is significantly related to urban poverty, but the sign is positive, implying that more arable land is associated with higher urban poverty. Presumably, urban areas in districts with a lot of agricultural land are small towns, so this variable may be picking up the effect of city size.

In addition, urban poverty is related to the percentage of the district covered with red-brown soils. It is difficult to explain this result; it may simply reflect a spurious correlation.<sup>24</sup> Urban poverty is also linked to the distance to the nearest town of more

than 10,000 inhabitants and to the distance to the nearest town of more than 100,000 inhabitants.

### **Selective Model of Urban Poverty**

A more selective model of the determinants of urban poverty was then developed using the same procedure as used for the rural model (Table 4.7). All soils apart from red-brown soils were successively eliminated by the stepwise regression procedure. Elevation continues to be a statistically significant predictor of poverty. However, in densely populated urban areas, it is not the percentage of land in different elevation classes that matters so much as the variation in elevation (as measured by the standard deviation of elevation). Because urban centers in the deltas and on the coast have very little variation in elevation, this variable shows that urban centers in midland and upland areas are poorer than those on the coast and the deltas. Finally, after some experimentation, the same three variables for distance to towns and cities were retained as in the urban model. However, in contrast to the rural model, distance from major and medium sized cities matters much more in urban areas. This indicates the importance of living close to major centers of demand; in particular, Hanoi and Ho Chi Minh City, the only two cities with populations of more

<sup>24</sup>Red-brown soils are not very common, accounting for an average of 4 percent of the area across districts. However, red-brown soils account for more than 50 percent of the area in 13 districts in four provinces: Gia Lai, Dak Lak, Binh Phuoc, and Dong Nai. In any regression analysis with more than 20 variables, it is likely that at least one coefficient will show a spurious correlation.

**Table 4.6 Inclusive model of the geographic determinants of urban poverty**

Characteristic	Coefficient	Robust standard error	z
Area between 251 and 500 m (%)	0.0001082	0.0002731	0.40
Area between 501 and 1,000 m (%)	-0.0000278	0.0002845	-0.10
Area between 1,001 and 1,500 m (%)	-0.000056	0.0005433	-0.10
Area over 1,500 m (%)	-0.0000753	0.00111	-0.07
Land cover (% of arable land)	0.0003578	0.0000948	3.77***
Land cover (% of bare and rocky land)	0.0003138	0.0005151	0.61
Land cover (% of national forest)	0.0001816	0.0001717	1.06
Land cover (% of plantation forest)	-0.0010288	0.0006689	-1.54
Slope (% of land with 4 to 8% slope)	0.0005427	0.0006581	0.82
Slope (% of land with 8 to 15% slope)	0.0007459	0.0009565	0.78
Slope (% of land with 15 to 30% slope)	0.0009161	0.0008757	1.05
Slope (% of land with over 30% slope)	0.0007792	0.0015636	0.50
Area with alluvial soils (%)	-0.0001124	0.0001845	-0.61
Area with alluvial glacial soils (%)	0.0001253	0.0002099	0.60
Area with alluvial acidic soils (%)	0.0000288	0.0002096	0.14
Area with acid sulfate soils (%)	0.0001295	0.000201	0.64
Area with salty soils (%)	0.0001363	0.0001762	0.77
Area with alluvial oxidized soils (%)	-0.0003464	0.0002077	-1.67*
Area with red-brown soils (%)	-0.0007689	0.0002382	-3.23***
Area with sandy soils (%)	0.0004235	0.0003504	1.21
Area with fluvial soils (%)	0.0001111	0.0002161	0.51
Area with Acrisol soils (%)	0.0003841	0.0002917	1.32
Area with other soils (%)	0.0001843	0.0002943	0.63
Area with rocky soils (%)	-0.0000639	0.0005357	-0.12
Area covered with water (%)	0.0001654	0.000528	0.31
Annual sunshine (days)	0.0000102	0.0000239	0.43
Annual precipitation (mm)	$-6.14 \times 10^{-6}$	0.0000206	-0.30
Distance from town with over 10,000 inhabitants (m)	$5.74 \times 10^{-7}$	$2.42 \times 10^{-7}$	2.37**
Distance from town with over 50,000 inhabitants (m)	$5.11 \times 10^{-7}$	$2.12 \times 10^{-7}$	2.41**
Distance from town with over 100,000 inhabitants (m)	$-1.40 \times 10^{-7}$	$1.79 \times 10^{-7}$	-0.78
Distance from town with over 250,000 inhabitants (m)	$-1.12 \times 10^{-10}$	$1.42 \times 10^{-7}$	-0.00
Distance from town with over 1 million inhabitants (m)	$1.17 \times 10^{-7}$	$1.02 \times 10^{-7}$	1.14
Constant	0.0992521	0.0687563	1.44
$\lambda$	0.8127389	0.0520617	15.61***

Notes: Spatial error model:  $N = 574$ ; variance ratio = 0.358; squared correlation = 0.289; log-likelihood = 865.43392;  $\sigma = 0.05$ ;  $R^2 = 0.3783$ . Wald test of  $\lambda = 0$ :  $\chi^2(1) = 243.706$  (0.000). Lagrange multiplier test of  $\lambda = 0$ :  $\chi^2(1) = 192.773$  (0.000). Acceptable range for  $\lambda$ :  $-1.900 < \lambda < 1.000$ . \* indicates significance at the 10 percent level, \*\* at the 5 percent level, and \*\*\* at the 1 percent level.

than a million, are more important to producers of nonagricultural products and services.

Note that the final version of the selective urban poverty model was estimated using the spatial error model, as rerunning the diagnostic tests for spatial dependence indicates that the spatial error model is now

preferred (Lagrange multipliers of 371.710 for the spatial error model against 293.1 for the spatial lag model) (Table 4.7).

It should be stressed, however, that the explanatory power of the selective urban model is rather modest: the six geographic variables included in the model explain just under 30 percent of the variation in urban



**Table 4.7 Selective model of the geographic determinants of urban poverty**

Characteristic	Coefficient	Robust standard error	z
Elevation of town (m)	0.0000132	0.0000229	0.58
Standard deviation of elevation	0.0001964	0.0000361	5.45***
Area with red-brown soils (%)	-0.0009041	0.0002049	-4.41***
Distance to town with over 100,000 inhabitants (m)	$2.39 \times 10^{-7}$	$1.71 \times 10^{-7}$	1.40
Distance to town with over 250,000 inhabitants (m)	$4.69 \times 10^{-8}$	$1.41 \times 10^{-7}$	0.33
Distance to town with over 1 million inhabitants (m)	$7.52 \times 10^{-8}$	$9.93 \times 10^{-8}$	0.76
Constant	0.1407774	0.019244	7.32***
$\lambda$	0.8023299	0.0504752	15.90***

Notes: Spatial error model:  $N = 573$ ; variance ratio = 0.339; squared correlation = 0.254; log-likelihood = 841.82739;  $\sigma = 0.05$ ;  $R^2 = 0.2917$ . Wald test of  $\lambda = 0$ :  $\chi^2(1) = 252.668$  (0.000). Lagrange multiplier test of  $\lambda = 0$ :  $\chi^2(1) = 371.710$  (0.000). Acceptable range for lambda:  $-1.900 < \lambda < 1.000$ . \*\*\* indicates significance at the 1 percent level.

poverty rates among districts. Furthermore, the explanatory power of the model does not improve very much when endogenous variables are included. Even after population, road density, and market density variables are added to the model, only 41 percent of urban poverty can be explained. Regional dummies are also relatively unimportant, increasing the fit ( $R^2$ ) of the selective model by 7 percentage points.

It is not surprising that agro-climatic conditions and market access matter much less to the industrial and service-related activities that drive productivity in urban areas. Here the clustering of industry, patterns of employment, and availability of complementary infrastructure will be crucial. A different, and probably more endogenous, set of variables would be needed to explain the geographic determinants of urban poverty.

## CHAPTER 5

---

### Spatial Variation in Determinants of Poverty

Chapter 4 described the results of an analysis of linkages between the spatial distribution of poverty and a number of agro-ecological and market access variables based on several different regression models. These models are “global” in the sense that they describe the relationship between poverty and geographic variables in Vietnam as a whole.

However, one would expect these relationships to vary over space. Indeed, an almost universal feature of spatial data is the variation in relationships over space, a phenomenon generally referred to as spatial nonstationarity or spatial drift. The problem of spatial nonstationarity is closely related to the problem of spatial dependence: the error terms of global regression models will show spatial autocorrelation if applied to data with spatially varying relationships because the global model can describe only universal relationships.

In this chapter, the results of an analysis of the spatial variation in relationships between poverty incidence on the one hand and a number of agro-ecological variables on the other are presented. We use a type of analysis called spatially weighted local regression analysis.

#### Model Description

The district-level estimates of  $P_0$  (the incidence of poverty) are defined as the dependent variable, and 14 variables from the database described in Chapter 2 were chosen as the independent variables. The 14 independent variables are listed in Table 5.1, and their spatial distributions are shown in Figure 5.1 (see color insert).

As described in Chapter 2, the variables were calculated at district level. The regression points were defined as the geographic center of each district of Vietnam, effectively fitting the defined data points. The regression points are those locations for which the local parameters are estimated and from which the other observations are weighted with decreasing weight the further away they are.

#### Results

A global regression model was first applied to the variables before the same regression model was “localized” by applying a geographically weighted regression. Table 5.1 presents the results of the global regression model. The  $R^2$  value of 0.74 indicates a reasonably good fit: 74 percent of the variance in district-level incidence of poverty is explained by the model. Almost all variables are significant (at the 1 percent confidence limit), and only market size (measured as payments of taxes from markets to the state) and total annual rainfall do not seem to be related to the dependent variable. Almost all the coefficients have the expected sign. For example, higher road density is expected to be negatively related to the incidence of poverty, meaning less poverty where the road density is higher. The only surprise is the positive relationship

**Table 5.1 Summary results of global model of rural poverty**

Variable	Coefficient	Standard error	<i>t</i> -statistic
Constant	124.4558	42.68905	2.9154***
Population density (%)	0.000597	0.000143	4.1799***
Natural forest (%)	-0.12724	0.023699	-5.3690***
Arable land area (%)	7.034238	2.069287	3.3993***
Bare land (%)	0.071751	0.041173	1.7426*
Markets per commune	-2.67395	0.695687	-3.8436***
Market payments to state	$-7.7 \times 10^{-7}$	$9.8 \times 10^{-7}$	-0.7855
Flat land (%)	-0.33163	0.021118	-15.7037***
Average distance to district town (km)	0.000665	0.000102	6.5066***
Main roads density (m/km <sup>2</sup> )	-0.01149	0.00259	-4.4370***
Minor roads density (m/km <sup>2</sup> )	-0.00324	0.001202	-2.69637**
Annual rainfall (mm)	0.000103	0.00114	0.0905
Average temperature (°C)	-1.10746	0.410026	-2.7009**
Annual sunshine duration (days)	-0.00253	0.00222	-1.1412
Average humidity (%)	-0.62011	0.502581	-1.2338

Note: \* indicates significance at the 10 percent level, \*\* at the 5 percent level, and \*\*\* at the 1 percent level.

between the percentage of arable land and the incidence of poverty because one would expect that in a largely agriculture-based economy, availability of arable land would have a positive influence on human welfare.

The question we would like to look at now is whether there are significant spatial variations in the relationships between the explanatory variables and the incidence of poverty. To investigate these questions, a local geographically weighted regression was applied to the variables. Table 5.2 shows the main diagnostic indicators of the global model compared to the respective results of the local model.

The local overall  $R^2$  (the average over the local regressions) is 0.95, indicating that 95 percent of the variance in district-level poverty incidence can be explained by the 14-variable local model. A local model will always fit the data at least as well as a global model, but the results suggest a large improvement in the fit, as measured by the  $R^2$ . Similarly, the residual sum of squares in the local model is about one-fifth that of the global model. Figure 5.2 shows the distribution of the residuals (the difference between the predicted poverty rate and the actual poverty rate). Both graphs display a rather

**Table 5.2 Summary results of global and local models**

Indicator	Result
Global $R^2$	0.74
Local overall $R^2$	0.95
Range in local $R^2$	0.83–0.99
Global residual sum of squares	44,908
Local residual sum of squares	8,778

normal distribution. Considering the results presented above, it is not surprising that the local model has a much narrower distribution than the global one, again indicating smaller errors and better fit in the local model.

However, there seem to be significant variations in the explanatory power of the local model over space. The map in Figure 5.3 (see color insert) shows the local  $R^2$  values obtained from the local regression analysis. Clearly, the values of  $R^2$  are everywhere above the global score (0.74), although the upland areas are generally better described by the model than the lowland and delta regions. This is not too surprising because one would expect agro-ecological and market access variables to have a stronger

**Table 5.3 Summary results of local parameter estimates**

Label	Minimum	Lower quartile	Median	Upper quartile	Maximum
Constant	-1,203.02	-53.1318	82.14617	395.9055	1,157.351822
Population density	-0.01106	0.000131	0.000232	0.00184	0.030141
Natural forest (%)	-0.64205	-0.08912	-0.031546	0.100038	0.295875
Arable land area (%)	-60.4885	-3.15322	3.700728	9.73332	74.51855
Bare land (%)	-2.01275	-0.05482	0.059338	0.48317	2.499394
Markets per commune (number)	-27.8087	-7.62955	-2.176135	-1.50498	1.758242
Market payments to state (million VND/6 mo.)	$-6.8 \times 10^{-5}$	$-7 \times 10^{-6}$	-0.000002	0	0.000023
Flat land (%)	-0.91052	-0.27857	-0.194937	-0.12296	0.201926
Average distance to district town (km)	-0.00074	0.000122	0.000512	0.000734	0.001523
Main roads density (m/km <sup>2</sup> )	-0.21916	-0.01899	-0.007749	-0.00303	0.020078
Minor roads density (m/km <sup>2</sup> )	-0.07833	-0.00879	-0.001636	-0.00053	0.010837
Annual rainfall (mm)	-0.04894	-0.01547	-0.004274	0.002622	0.037291
Average temperature (°C)	-41.5685	-5.20404	-2.874492	1.298637	24.52939
Annual sunshine duration (days)	-0.08096	-0.0475	-0.011726	0.005215	0.187103
Average humidity (%)	-16.7322	-3.18053	-0.280914	2.020089	15.454768

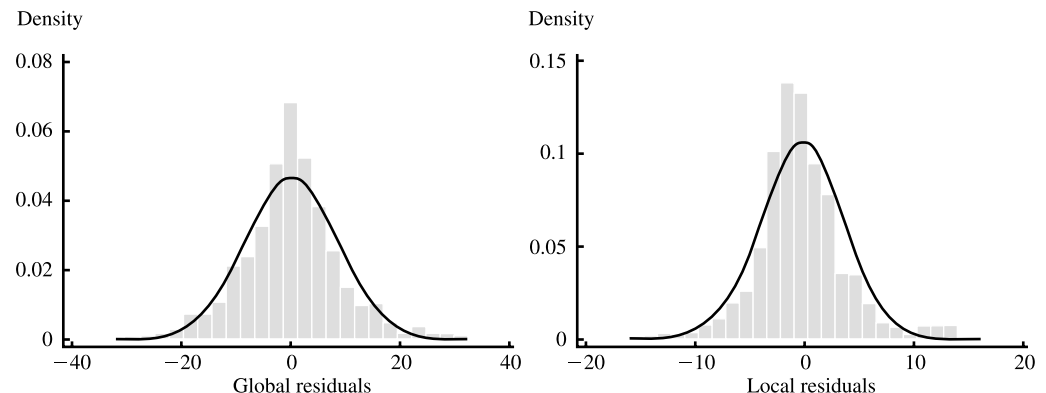
influence on human welfare in environmentally more difficult areas than in more accessible and less mountainous areas.

We now turn our attention to the exploration of spatial variations in the variable relationships.

Table 5.4 summarizes the local coefficients estimated by the local model, and Figure 5.4 (see color insert) visualizes the spatial distribution of those local regression coefficients for each of the 14 explanatory variables. Clearly, a great deal of variation in relationships over space can be identified for almost all of the parameters, and many of them have rather large ranges, often even including changes of signs.

Significance tests based on Monte Carlo simulations are used to see whether the observed spatial variations in parameter estimates are caused by random variations or whether they reflect true spatial differences. The tests indicate that all variables except for “average distance to district town” have statistically significant spatial variation with a high level of confidence (see Table 5.4).

Although the interpretation of the variation in relationships over space can be rather tricky for some variables, it seems more intuitive for others. For example, the average number of markets per commune appears to be much more negatively related to the incidence of poverty in poorer, remote upland

**Figure 5.2 Distribution of residuals in the global and local models**

**Table 5.4 Significance of spatial variations in parameter estimates**

Parameter	<i>P</i> -value
Population density	0.00***
Natural forest (%)	0.00***
Arable land area (%)	0.00***
Bare land (%)	0.01***
Markets per commune	0.00***
Market payments to state	0.03**
Flat land (%)	0.01***
Average distance to district town	0.84
Main roads density	0.00***
Minor roads density	0.00***
Annual rainfall	0.00***
Average temperature	0.00***
Annual sunshine duration	0.00***
Average humidity	0.00***

Note: \*\* indicates significance at the 5 percent level and \*\*\* at the 1 percent level.

areas than in better-off lowland areas. A similar pattern can be identified for road densities.

As for the relationship of forest cover to poverty, it appears that availability of natural forest generally tends to have more negative relationship to poverty in remote upland areas where there is generally a smaller percentage of forest cover (e.g., the Northern Uplands) compared to other upland areas (e.g., the Central Highlands) (see Figs. 3.6 and 5.1 in color insert). The positive relationship between natural forest cover and poverty in some lowland and less remote areas indicates that poorer districts tend to have higher percentages of forest cover than richer districts in the same region.

A somewhat surprising outcome can be seen in the spatial patterns of the relationship between population density and poverty: although one might expect population density to be negatively correlated to poverty (i.e., higher population density relates to lower poverty rates) primarily in remote areas, Figure 5.4 appears to indicate

the opposite. Apparently, in several remote areas, despite relatively low population densities, population pressure appears to negatively affect human welfare in these remote upland areas. This could be explained by the low “carrying capacity” of these upland areas, caused mainly by their low agricultural productivity.

Though the climatic variables show great variation in their relationships to poverty incidence, the interpretation is less straightforward and often not necessarily intuitive. Nevertheless, some patterns appear obvious: Although globally not significantly related to poverty, annual rainfall appears in large parts of Vietnam to be negatively correlated to poverty (i.e., more rainfall relates to lower poverty rates), it seems that the central coast does not benefit from higher annual rainfall. This might well be related to the area frequently being affected by typhoon-induced floods causing great damage in the region almost on an annual basis.

Spatial patterns in relationships between poverty and agro-ecological and market access variables show great variations over space. Some variables might be important positive contributors to some socioeconomic conditions in one area, but the same variables could have no, or even a negative, effect on the same socioeconomic situation in other areas. A more localized analysis approach helps to reveal local particularities not necessarily obvious otherwise. Such information can be crucial for informed spatially disaggregated pro-poor policy making. In fact, conclusions drawn from “global” results applied to “local” conditions might in some cases even be misleading. Future research in this direction, for instance localized analysis of urban and rural poverty in relation to agro-ecological and market access variables separately, would therefore promise valuable new insights into local aspects of the poverty–environment nexus.

## CHAPTER 6

---

### Summary and Conclusions

**T**his concluding chapter is divided into four sections. The first provides a brief descriptive summary of the objectives, methods, and results of this report, and the second describes a number of general conclusions that can be drawn. Although the main purpose of this study was to generate information on the patterns of poverty and inequality rather than to address specific policy issues, the third section explores some of the implications of the findings for policy and programs. Finally, because the report raises as many questions as it answers, the last part describes some avenues for future research related to poverty and inequality in Vietnam.

#### Summary

##### Background and Methods

Information on the spatial distribution of poverty and inequality is useful because it assists policymakers and program designers by shedding light on the causes of poverty and facilitating efforts to target poverty alleviation programs on the poorest regions. Information on the spatial patterns of poverty is particularly important in Vietnam because of the large regional disparities within the country and the government's strong commitment to the goal of reducing poverty and eliminating hunger.

Various household surveys carried out by the General Statistics Office (GSO) provide information on the incidence of poverty at the regional and provincial levels, but these surveys cannot generate district- or commune-level poverty estimates. The Ministry of Labor, Invalids, and Social Affairs (MOLISA) produces a list of the poorest communes based on data collected administratively from its officers in the field, but the criteria and methods differ across provinces.

The objectives of this study were (1) to explore the spatial distribution of poverty and inequality in Vietnam using small-area estimation methods, (2) to study the effect of agro-climatic variables and market access on poverty at the district level, and (3) to demonstrate the potential for new methods, including the small-area estimation method, to generate information useful for policy makers and program designers in Vietnam.

The small-area estimation method combines household survey data and census data to estimate a variable of interest (often poverty) for small areas such as districts. First, the 1997–98 Vietnam Living Standards Survey data are used to estimate an equation describing the relationship between per capita expenditure and various household characteristics. The equation is then applied to data on those same household characteristics from the 1999 Population and Housing Census, generating poverty estimates for each household in the Census. These results

are then aggregated to generate estimates of poverty and inequality at the commune, district, and province level.

The study examines the geographic determinants of poverty with spatial regression analysis. The dependent variable is the district-level poverty rate, and the explanatory variables include variables representing topography, soil, climate, and market access.

### Spatial Patterns in Poverty and Inequality

In the first stage of the analysis, an econometric model of per capita expenditure is estimated for urban and rural households using the VLSS. As expected, per capita expenditure is significantly related to household size and composition, education of the head of household and the spouse, housing characteristics, source of water, electrification, type of toilet, source of water, and region of residence. In general, the models explain a little more than half of the variation in per capita expenditure in urban and rural areas.

With regard to the spatial patterns in the incidence of poverty, the findings can be summarized as follows:

- Poverty rates ( $P_0$ ) are highest in the Northeast and Northwest along the border with China and the Lao P.D.R., in the interior of the central coast, and in the northern part of the Central Highlands.
- Poverty rates are intermediate in the two main deltas of Vietnam, the Red River Delta and the Mekong Delta.
- Poverty rates are the lowest in the large urban areas, particularly Hanoi and Ho Chi Minh City, and in the Southeast region.
- Urban poverty rates are consistently much lower than rural poverty rates.
- The map of commune-level poverty reveals the effect of mountains and even highways on poverty rates.
- There is only a very weak correlation between these district-level poverty estimates and those estimated by MOLISA.

The confidence intervals for the province- and district-level poverty estimates are reasonable: half the provinces have confidence intervals between  $\pm 4.2$  and  $\pm 5.8$  percentage points, whereas half the districts have intervals between  $\pm 4.4$  and  $\pm 6.9$  percentage points. However, the confidence intervals for poverty at the commune level are higher, indicating less reliable estimates. Half the communes have confidence intervals between  $\pm 6.6$  and  $\pm 10$ .

Two other poverty measures, the depth of poverty ( $P_1$ ) and the severity of poverty ( $P_2$ ), were estimated at the district level. These two measures were highly correlated with the incidence of poverty ( $P_0$ ), resulting in very similar poverty maps.

The map of the density of poverty (the number of poor people per unit of area) reveals that the density of poverty is greatest where the incidence of poverty is lowest. The regions with the highest poverty rates, the Northeast, Northwest, and Central Highlands, are so sparsely populated that the number of poor people living in them is relatively small. In contrast, the densely populated cities and the deltas account for a greater absolute number of poor people despite their lower poverty rates.

This study also generated district-level estimates of three measures of inequality in per capita expenditure: the Gini coefficient, the Theil L index, and the Theil T index. We can summarize the results as follows:

- The three measures of inequality are very highly correlated.
- Inequality is greatest in the large urban areas and in parts of the Northeast, Northwest, and Central Highlands.
- Inequality was the lowest in the Red River Delta, followed by the Mekong River Delta.
- More than three-quarters of the inequality is within provinces rather than between provinces.
- About two-thirds of the inequality is within districts rather than between districts.

Examining the relationships among poverty, inequality, per capita expenditure, and the degree of urbanization at the district level, the study found that:

- The district-level poverty rate and average per capita expenditure in the district are highly correlated, with per capita expenditure explaining about 96 percent of the variation in inequality across districts.
- In general, higher per capita expenditure is associated with higher inequality, but some poorer districts also have very high levels of inequality.
- Inequality is highest in districts with very low or very high poverty rates, although the correlation is not very high.
- As the share of the population living in urban areas rises, the poverty rate declines.
- As the share of the population living in urban areas rises, the level of inequality rises up to a point, after which further urbanization is associated with lower inequality. In other words, the districts with the highest levels of inequality are those with both urban and rural populations.

### **Geographic Determinants of Poverty**

This analysis explored the geographic determinants of poverty using spatial regression analysis. The dependent variable is the district-level incidence of poverty ( $P_0$ ), and the explanatory variables include a wide range of GIS variables including elevation, slope, soil type, land cover, rainfall, sunshine, and distance to towns and cities. Separate models are used to estimate urban and rural poverty.

Statistical tests on the rural model indicate that there is spatial correlation in the error terms, meaning that some geographic variables not in the model are causing the error terms in neighboring districts to be correlated. We compensate for this by running generalized least squares with a spatial weighting matrix. An inclusive model with

32 explanatory variables explains about 80 percent of the variation in district-level poverty rates. The following factors are positively linked to rural poverty: bare and rocky land cover, steep slopes, acid sulfate soils, sandy soils, salinated soils, and distance to a town of at least 10,000 inhabitants. Elevation, annual rainfall, and annual hours of sunshine do not have statistically significant effects. After collapsing categories and removing insignificant variables, we get a more selective model of rural poverty in which 12 explanatory variables explain 74 percent of the variation in district level poverty. In this version, distance to cities of at least 100,000 inhabitants is also significant.

Tests of the spatial correlations in the model of urban poverty were somewhat ambiguous but suggested that we should again use the spatial error model. The inclusive model with 32 explanatory variables explains just 38 percent of the variation in urban poverty. This implies that urban poverty is much less affected by agro-climatic conditions and market access than rural poverty. Urban poverty is positively associated with arable land, distance to towns of at least 10,000 inhabitants, and distance to towns of at least 100,000 but negatively associated with red-brown soils. In the selective model, six variables explain 29 percent of the variation in urban poverty.

### **Spatial Variation in the Determinants of Poverty**

The nature of how agro-climatic and market access variables are related to poverty varies over space. This study applied a geographically weighted local regression analysis to explore the geographic variation in relationships of 14 explanatory variables to the incidence of poverty.

The results of the analysis suggest two things. First, there are significant variations in the way that individual explanatory variables are related to poverty. The association between market access and poverty, and between access to natural forests and poverty, is strongest in comparatively disadvantaged



areas. In most parts of Vietnam, higher rainfall is associated with lower poverty, but in some areas, the reverse is true. This appears to reflect vulnerability to environmental stress, such as flooding or typhoons. Second, a model that allows for spatial variations in relationships better describes the complex relationship between poverty and the environment. The results of the analysis suggest that the local measures of goodness of fit are everywhere higher than the fit achieved by the global model but that there are still great differences over space in how the model can replicate the data. Generally, areas with a more difficult terrain achieved a better fit in the model, suggesting that agro-ecological and market access conditions have a stronger influence on human welfare than in areas where environmental conditions are less difficult.

## Conclusions

The district poverty rates estimated in this study differ significantly from the poverty rates estimated by MOLISA. Although MOLISA uses a different welfare indicator, a different poverty line, a different method of adjusting for the local cost of living, and measures poverty at the household rather than the individual level, it is still surprising that there is only a weak correlation between the poverty rates estimated by the two methods.

Poverty rates across districts and communes vary widely. One of the striking aspects of the poverty maps generated by this study is the wide variation in poverty rates. In some districts, particularly remote districts in the upland areas, over 90 percent of the population lives below the poverty line. In others, particularly in or near the large urban centers, less than 5 percent of the population is poor.

In spite of the wide variation in poverty rates across the country, the level of inequality is relatively low. One might expect a country with such wide variations in poverty to have a high degree of inequality, but

the level of inequality in Vietnam is relatively low by international standards. One possible explanation is that the poorest areas tend to be sparsely populated, so they do not greatly affect national inequality figures. The bulk of the rural population in Vietnam lives in the Mekong Delta and the Red River Delta, where inequality is relatively low. Even more surprising is the fact that variations in average per capita expenditure across the districts accounts for just one-third of the total inequality in the country. Inequality within districts accounts for two-thirds of the total.

High levels of inequality are not found only in urban and commercial farming areas. It is commonly believed in Vietnam (and in other developing countries) that inequality is primarily associated with urban areas and rural areas characterized by commercial agriculture. This is based on the idea that inequality is the byproduct of economic growth, as some households take advantage of new market opportunities and earn incomes much higher than average. Our findings confirm that inequality is greater in urban areas than rural areas, but we also find that inequality can be quite high in rural areas, even in areas characterized by sparse population and semisubsistence farming.

Differences in poverty across districts are mainly determined by differences in average per capita expenditure, not the degree of inequality. Ninety-six percent of the variation in district-level poverty rates can be explained by differences in average per capita expenditure, with differences in inequality accounting for less than three percent. The explanation is that inequality does not vary much from one district to another.

Most poor people live in the less poor areas. The density of poor people is lowest in areas with the highest poverty rate (such as the rural upland areas), and the poverty density is highest in areas with low poverty rates (such as cities and rural deltas). The absolute number of poor people that live in areas with high poverty rates is relatively low because the population density in these areas is also low. By contrast, most of the

rural poor live in the Mekong Delta and the Red River Delta. Although these areas have relatively low poverty rates compared to other rural areas, the population density ensures that most of the poor live in the two deltas.

Most of the variation in district-level rural poverty can be explained by agro-climatic factors and market access. Although it is not surprising that agro-climatic factors and market access explain some of the variation in district-level rural poverty, it is somewhat surprising that they explain three-quarters of the variation. In contrast, less than two-fifths of the variation in urban poverty can be explained by these factors.

### **Implications for Policy and Programs**

The main objective of this study is to examine spatial patterns in poverty and inequality, with the idea that this information is useful for targeting poverty alleviation programs. The study was not designed to assess specific policy options for reducing poverty. The results do, however, provide some indirect implications for policy and programs. In this section, we discuss some of those implications.

#### **Where Are the Poor?**

The most obvious application of the results presented in this report is in improving information on the spatial distribution of poverty for the purpose of targeting poverty alleviation programs. Not only do the results provide information on the distribution of poverty in Vietnam, but they also provide information on the accuracy of these estimates. In addition, by generating information on alternative poverty measures, they allow program designers to target assistance on districts with the greatest depth or severity of poverty.

#### **Assistance to Poor Areas or Poor People?**

If most poor people live in less poor areas, what are the implications for targeting pov-

erty alleviation programs? In particular, should poverty alleviation programs concentrate their efforts on areas with the greatest poverty density? The answers depend on the type of poverty alleviation program, as discussed below.

Some programs are relatively untargeted and lift the income of all households in an area. Examples might be better roads, better health care, and financial support to local government. Assuming the program has a fixed cost per inhabitant, the program will have a greater effect on poverty if it is concentrated on poor areas. In these areas, a higher percentage of the population is poor, so a higher percentage of the beneficiaries will be poor. In this way, the government achieves more poverty reduction per dollar spent. This is certainly true if the goal is to reduce the depth of poverty ( $P_1$ ), and it is probably true if the goal is to reduce the incidence of poverty ( $P_0$ ).

Other programs are targeted to poor households (e.g., income transfers, food for work, or social service fee exemptions). If the goal is to provide the same level of assistance to each poor person, the program should spend more overall in areas with many poor people (such as the deltas in Vietnam) but more money per inhabitant in areas with high poverty rates (such as the Northern Uplands and the Central Highlands).

Of course, these guidelines assume the cost of providing the program is constant in per capita terms, implying that the cost is not affected by population density. Some programs, such as electrification and extension, will cost more in per capita terms in low-density areas. Other programs, particularly land-intensive programs such as roads and parks, may be more expensive in a high-density area.

#### **Does Geography Make Upland Development Impossible?**

The analysis of the geographic determinants of poverty reveals that three-quarters of the variation in rural poverty at the district level can be explained by a small number of agro-

climatic and market access variables. This finding is somewhat troubling because it is not possible to design policy interventions that directly influence the agro-climatic variables. These results might be interpreted as saying that those living in districts with steep slopes and poor soils are caught in spatial poverty traps from which it is difficult to escape.

We are less pessimistic about these findings. First, market access can be influenced by public investment and policy. Although the government cannot reduce the actual distance to cities, it can reduce travel time and travel cost, which are probably the relevant variables. Of course, roads will also allow goods produced more cheaply elsewhere (such as rice) to enter the region and compete with local production. But trade theory suggests that the aggregate impact on the region will be positive, and the results presented here indicate it will even be positive in terms of reducing poverty.

In addition, geography is a limiting factor in poverty reduction only to the extent that people are not able to migrate. To the extent that migrants are able to raise their living standards without negatively affecting others, migration can be an effective tool to reduce poverty. The implication is that the government should not exclude migration as a possible development strategy, particularly for districts that are severely constrained by agro-ecological factors. Relaxing some of the restrictions on migration would allow people from agro-climatically constrained areas to raise their incomes and reduce poverty. Although migrants from rural areas to the cities tend to be initially poorer than their urban neighbors, thus contributing to a more visible increase in the number of urban poor, the relevant question is whether the standard of living of the migrants is better than it would be if they had not migrated.

Finally, it is important to avoid the idea that geography will prevent any development in disadvantaged areas. Other studies have shown that economic growth and pov-

erty reduction have occurred even in disadvantaged regions such as the Northern Uplands (Poverty Working Group 1999). The fact that agro-climatic factors are good predictors of poverty rates across districts at one point in time does not mean that they are good predictors of poverty over time for a given district.

### **Growth versus Equity**

In Vietnam, as elsewhere, there is a debate between those who support policies and programs to reduce poverty through direct assistance to poor people and those who support policies and programs to increase economic growth as a strategy to raise the poor out of poverty. This study finds that almost all (96 percent) of the variation in district-level poverty rates can be explained by differences in district-level average per capita expenditure. Certainly, it is possible to reduce district-level poverty by reducing inequality, but in practice this is not what distinguishes high- and low-poverty districts in Vietnam. If this cross-sectional pattern reflects the changes that occur over time, then the implication is that poverty reduction occurs largely as a result of broad-based economic growth rather than improvements in income distribution. Indeed, this is consistent with the results of a comparison of the 1992–93 and 1997–98 Vietnam Living Standards Surveys (see Poverty Working Group 1999). This finding highlights the importance of policies and programs that promote household income growth as a strategy to reduce poverty.

### **Implications for Future Research**

Poverty at the household level can be explained fairly well on the basis of simple household characteristics. Over half of the variation in per capita expenditure can be explained using 17 household characteristics from the Census questionnaire. These variables cover household size and composition, occupation, education, housing char-

acteristics, water and toilet facilities, ownership of consumer durables, and region of residence. A questionnaire focused on the characteristics that distinguish poor from nonpoor households should be able to predict expenditures even better. This suggests the potential for developing a short survey (or a set of indicators to be included in larger surveys) that would focus on household characteristics proven to be associated with expenditure or income. This could be used for poverty monitoring, project evaluations, or household-level targeting. Some work on this topic has been done in Vietnam (see Minot and Baulch 2002a; Baulch 2002), but more work is needed to identify and build consensus around the best predictors of poverty and verify that the targeting based on these predictors would be reasonably accurate.

Further research is also needed to evaluate the discrepancies between our poverty estimates and those of MOLISA. The estimates provided in this study may be flawed if, in some areas, poverty is not well predicted on the basis of household characteristics. However, the MOLISA estimates may also be flawed because of inconsistencies in their definition of poverty or poverty-monitoring procedures (Conway 2001). One approach would be to study in more depth a number of districts in which the two methods give very different poverty estimates.

Small-area estimation is a valuable tool for understanding the spatial distribution of poverty and inequality. The results presented in this report suggest that there is considerable potential for using small-area estimation methods and census data to obtain a better understanding of the spatial patterns in poverty and inequality. Census data provide the level of disaggregation that will be increasingly necessary for spatially disaggregated policy analysis and decentralization.

However, small-area estimation cannot easily be used to update poverty maps. Although small-area estimation is valuable for

generating poverty maps and other information about the spatial distribution of poverty and inequality, it probably cannot be used to generate district and commune poverty estimates for all of Vietnam on an annual basis. If the analysis uses census data in the second stage, it can be updated only every 10 years. Data from the agricultural census could be used to update the estimates of rural poverty every 5 years. Annual household surveys, such as those carried out by GSO, can only help update the prediction equation, not the poverty estimates themselves. To update district-level poverty estimates would require a simple survey (with a questionnaire similar to that of a census), but with a very large sample, perhaps 600,000 households.

Small-area estimation can also be applied to the study of the spatial distribution of nutrition, commercial agriculture, or any other variable that can be predicted based on household characteristics. Although this report applies small-area estimation methods to study the spatial patterns in poverty and inequality, the method could be used to explore spatial patterns in other variables of interest. For example, if caloric malnutrition or micronutrient deficiencies can be predicted using household characteristics in a nutrition survey, the results could be applied to the census data to produce detailed information on the spatial distribution of those problems. Similarly, other variables such as the degree of income diversification, vulnerability to weather-related shocks, or involvement in commercial agricultural production could be mapped in a similar way if they can be predicted with at least moderate accuracy by household characteristics in the census data. Another possible application is to use small area estimation methods to examine poverty and inequality among groups of households that are too small to be studied with conventional household surveys, such as the disabled, specific ethnic groups, widows, or marginalized occupational groups.

## APPENDIX A

---

### Comparison of Results Using Different Analysis Methods

#### Introduction

**A**s described in Chapter 2, the small-area estimation analysis that generates estimates of poverty and inequality was carried out using a program written by the authors in Stata software. The estimates of the incidence of poverty ( $P_0$ ) and the standard errors of  $P_0$  were based on the formulas described by Hentschel et al. (2000). These formulas assume the error term in the first-stage regression model of per capita expenditure is homoskedastic and has no spatial autocorrelation. The estimation of other poverty measures ( $P_1$  and  $P_2$ ) and the three measures of inequality (the Gini, Theil L, and Theil T) were calculated by running the Hentschel procedure 100 times with different “poverty lines,” constructing a numerical cumulative distribution function for per capita expenditure, and then using this function to estimate the poverty and inequality measures. The Stata program uses Huber/White/sandwich estimators for the standard errors, which, although they do not explicitly model heteroskedasticity and spatial autocorrelation, do take into account the sample structure (including clustering) and are robust to heteroskedasticity. The Stata program is small, flexible, and runs quickly, but it does have two important limitations: it does not explicitly take heteroskedasticity and location effects into account in the first-stage regression analysis, and it does not calculate the standard errors for  $P_1$ ,  $P_2$ , and the measures of inequality.

Elbers, Lanjouw, and Lanjouw (2003) propose an alternative approach to generating small-area estimates. In this approach, the first-stage regression model is estimated with generalized least squares, taking into account heteroskedasticity and location effects<sup>25</sup> and then calculating the poverty and inequality measures and their standard errors using simulation methods. Although more computationally intensive, this approach allows estimation of all the major indicators of poverty and inequality as well as the standard errors of all the estimates. With support from the World Bank, Gabriel Demombynes has written a program in SAS to implement the approach developed by Elbers, Lanjouw, and Lanjouw (2003).

#### Method

This raises the question as to whether the Stata program produces substantially different results compared to the more comprehensive SAS program. In order to answer this question, we used

---

<sup>25</sup>Location effects refer to the fact that the error terms in the regression model are likely to be correlated within each cluster of households in the sample because of cluster-specific factors not included in the model such as soil type and access to markets. Location effects are one type of spatial autocorrelation.

the SAS program to run the first-stage regression analysis of the 1997–98 Vietnam Living Standards Survey and generate poverty and inequality estimates from a subset of the 1999 Population and Housing Census. In particular, we examined the district-level estimates of three measures of poverty ( $P_0$ ,  $P_1$ , and  $P_2$ ) and three measures of inequality (Gini, Theil L, and Theil T) using data from the rural areas of three provinces: Lai Chau, Ha Tinh, and Ba Ria–Vung Tau. These three provinces were selected to represent the range of different levels of development within Vietnam. Lai Chau is a remote, mountainous, and largely agricultural province in the extreme northwest of the country and has the highest poverty rate in Vietnam, according to our estimates (80 percent). Ha Tinh is a small coastal province in the North Central Coast region with an intermediate poverty rate (45 percent). Ba Ria–Vung Tau is a more urbanized coastal province in the Southeast region, benefiting from proximity to Ho Chi Minh City, local tourism, and substantial employment in the oil industry. The poverty rate in Ba Ria–Vung Tau is the third lowest in the country (10 percent). We focus on rural areas to simplify the analysis and because geographically targeted programs in Vietnam concentrate on rural areas because of the higher poverty rates there.

The SAS program can be run with or without adjustments for heteroskedasticity and location effects. In this analysis, we make two comparisons. First, we compare the results of the Stata program to those of the SAS program without adjustment for heteroskedasticity and location effects. Assuming there are no programming errors, these two approaches differ mainly in the way that the standard errors are calculated, though this may have small effects on the poverty and inequality estimates themselves.<sup>26</sup> The second comparison is between

the results of the Stata program and those of the SAS program with adjustment for location effects and heteroskedasticity. The difference between the sets of results will reflect both the different method of calculating standard errors and the impact of explicitly modeling heteroskedasticity and location effects.

## Results

### Comparison of First-Stage Results

In the first stage, regression analysis is used to “predict” per capita expenditure in rural areas, the Stata program and the SAS program without adjustments generate coefficients that are identical down to the fifth decimal (the precision used in the display of the SAS results). The standard errors, however, are substantially larger in the Stata version, presumably because the Stata program generates robust standard errors rather than ordinary standard errors (assuming normal and identical, independently distributed error terms).

The SAS program with adjustments generates somewhat different coefficients in the first-stage regression model (the average absolute value of the difference is 32 percent of the Stata coefficient). Somewhat surprisingly, the standard errors in the Stata model are generally larger than those in the SAS model with adjustments. More specifically, 75 percent of the coefficients have larger standard errors in the Stata model. All the coefficients that are statistically significant at the 5 percent level in the Stata model are also statistically significant in the SAS model with adjustments. One coefficient (for ethnic minorities) was significant in the SAS model with adjustments but not in the Stata model. Thus, it appears that using Stata with robust Huber/White/sandwich standard

<sup>26</sup>Recall that in equation 2, the expected probability that a household is poor is partly a function of  $\sigma$ , the standard error of the regression. If  $\sigma$  is underestimated, the probability that the household is poor will be biased away from 50 percent, and the degree of inequality will be overestimated.

errors does not exaggerate the statistical significance of the explanatory variables compared to a generalized least-squares model that incorporates heteroskedasticity and location effects.

### Comparison of Second-Stage Results

In the second stage, the estimated regression equation is applied to the subset of the Census data. The district-level estimates for the incidence of rural poverty ( $P_0$ ) produced by the Stata program are quite similar to those produced by the SAS model without adjustments. Table A.1 shows that, across the 26 districts in the three selected provinces, the difference ranges from 0.000 to 0.006, with most being less than 0.003.<sup>27</sup> Across the 26 districts, the mean difference in the estimates of  $P_0$  is less than 0.0005,<sup>28</sup> and the mean absolute difference is 0.003. The mean absolute difference (0.003) represents less than 1 percent of the average point estimate of  $P_0$  (0.510). Because the first-stage regression coefficients are identical in the Stata program and the SAS program without adjustment, it is not surprising that the poverty estimates are so close. These small differences in poverty estimates must be caused by differences in the way standard errors are calculated (ordinary vs. robust) and perhaps differences in the way the poverty estimates are generated (simulation vs. formula).

Table A.1 also compares the Stata results with the SAS results with adjustment. In this case, the mean difference is about 1 percentage point (−0.012), and the mean absolute difference is about 3 percentage points (0.032). This mean absolute difference is about 6 percent of the average estimate of  $P_0$ . Clearly, taking heteroskedasticity and location effects into account has a larger effect on the poverty estimates than switching from the analytic solution of Hentschel et al.

(2000) to the simulation-based approach of Elbers, Lanjouw, and Lanjouw (2003).

Table A.2 compares the district-level rural Gini coefficients generated by the Stata program and the SAS program. All the estimated Gini coefficients are in the range 0.21 to 0.27. The Stata estimates of the Gini coefficient tended to be slightly lower than the SAS estimates without adjustment but slightly higher than the SAS estimates without adjustment. In other words, the Stata program overestimated inequality somewhat, producing estimates 0.005 higher on average than the SAS program with adjustments or 2 percent of the average Gini estimate by the SAS program with adjustment. In absolute value, the Stata estimates of the Gini coefficient differed by an average of 0.011 compared to those of the SAS program with adjustment, representing less than 5 percent of the average Gini estimate by the SAS program with adjustment.

Table A.3 shows the comparison of the three methods in estimating the standard error of  $P_0$ , the other two measures of poverty ( $P_1$  and  $P_2$ ), and the other two measures of inequality (Theil L and Theil T). To save space, we suppress the results for each district and report only the means over the 26 districts. As expected, the standard errors of  $P_0$  generated by the Stata program are larger than those produced by the SAS program without adjustment. This is because the Stata program calculates robust standard errors in the first-stage regression, which are used in calculating the standard errors of the poverty estimates (see equation 5). However, the standard errors produced by the Stata program are smaller than those produced by the SAS program with adjustments. On average, the standard errors from the Stata program are about 40 smaller than those produced by the SAS program with adjustment. In other words, the Stata program, by

<sup>27</sup>For example, the difference between a headcount poverty rate of 45.0 percent and 45.3 percent is 0.003.

<sup>28</sup>Although not shown in the table, the mean difference was actually 0.0001.

**Table A.1 Comparison of  $P_0$  estimates using different methods**

District	Stata (1)	SAS without adjustment (2)	SAS with adjustment (3)	Difference (1) – (2)	Difference (1) – (3)
Lai Chau province					
30101	0.592	0.594	0.667	–0.002	–0.076
30105	0.948	0.950	0.973	–0.002	–0.024
30107	0.894	0.899	0.930	–0.005	–0.036
30109	0.943	0.947	0.972	–0.004	–0.030
30111	0.918	0.922	0.943	–0.003	–0.025
30113	0.937	0.941	0.968	–0.004	–0.031
30115	0.922	0.927	0.952	–0.005	–0.031
30117	0.710	0.714	0.782	–0.003	–0.071
30119	0.935	0.939	0.957	–0.003	–0.021
Ha Tinh province					
40501	0.347	0.344	0.362	0.003	–0.014
40503	0.424	0.421	0.430	0.003	–0.006
40505	0.461	0.460	0.495	0.001	–0.034
40507	0.411	0.409	0.423	0.002	–0.011
40509	0.453	0.451	0.474	0.002	–0.021
40511	0.504	0.504	0.522	0.000	–0.018
40513	0.487	0.486	0.512	0.001	–0.025
40515	0.487	0.486	0.517	0.001	–0.029
40517	0.550	0.551	0.570	–0.001	–0.020
40519	0.514	0.514	0.552	0.000	–0.038
Ba Ria–Vung Tau province					
71701	0.139	0.135	0.096	0.004	0.043
71703	0.086	0.082	0.051	0.003	0.035
71705	0.107	0.103	0.071	0.004	0.037
71707	0.171	0.167	0.117	0.004	0.054
71709	0.130	0.124	0.091	0.006	0.039
71711	0.127	0.124	0.089	0.003	0.038
71713	0.067	0.068	0.052	–0.001	0.015
Mean value	0.510	0.510	0.522	0.000	–0.012
Mean absolute value				0.003	0.032

Source: Analysis of data from 26 districts in three provinces to calculate poverty and inequality measures using three methods: the Stata program used in this report, an SAS program developed by the World Bank with no adjustments for heteroskedasticity and location effects, and the same SAS program with adjustments for heteroskedasticity and location effects.

not taking into account heterogeneity and location effects, underestimates the margin of error of the estimates of the incidence of poverty.

Looking at the two measures of poverty in Table A.3, it appears that the Stata program underestimates the poverty gap ( $P_1$ ) and the squared poverty gap ( $P_2$ ). The bias is about 7–8 percent of the SAS value with adjustment for both  $P_1$  and  $P_2$ . The mean absolute difference between the district-level

estimates of the Stata program and those of the SAS program with adjustments is 9 percent of the SAS estimate in the case of both  $P_1$  and  $P_2$ .

In contrast, the Stata program overestimates inequality as measured by the Theil L and Theil T indexes. Compared to the SAS estimate with adjustment, the average upward biases of the Stata estimates of Theil L and Theil T are 23 and 17 percent, respectively. The mean absolute difference between



**Table A.2 Comparison of Gini coefficient estimates using different methods**

District	Stata (1)	SAS without adjustment (2)	SAS with adjustment (3)	Difference (1) – (2)	Difference (1) – (3)
Lai Chau province					
30101	0.262	0.269	0.243	–0.006	0.019
30105	0.239	0.245	0.221	–0.006	0.018
30107	0.251	0.263	0.242	–0.012	0.009
30109	0.225	0.234	0.210	–0.009	0.015
30111	0.254	0.264	0.246	–0.010	0.008
30113	0.228	0.235	0.211	–0.007	0.018
30115	0.224	0.241	0.219	–0.017	0.005
30117	0.264	0.277	0.256	–0.013	0.008
30119	0.245	0.253	0.230	–0.007	0.015
Ha Tinh province					
40501	0.273	0.253	0.230	0.020	0.043
40503	0.235	0.287	0.270	–0.053	–0.035
40505	0.234	0.231	0.210	0.003	0.024
40507	0.225	0.238	0.222	–0.013	0.003
40509	0.225	0.234	0.217	–0.009	0.008
40511	0.216	0.233	0.216	–0.016	0.000
40513	0.227	0.231	0.215	–0.003	0.013
40515	0.228	0.238	0.222	–0.010	0.007
40517	0.232	0.232	0.216	0.000	0.016
40519	0.223	0.240	0.222	–0.016	0.001
Ba Ria–Vung Tau province					
71701	0.234	0.233	0.218	0.001	0.017
71703	0.247	0.246	0.232	0.001	0.015
71705	0.243	0.258	0.253	–0.015	–0.010
71707	0.246	0.253	0.253	–0.007	–0.007
71709	0.257	0.257	0.254	0.000	0.003
71711	0.256	0.274	0.275	–0.017	–0.019
71713	0.265	0.269	0.267	–0.004	–0.002
Mean value	0.241	0.251	0.235	–0.010	0.005
Mean absolute value				0.010	0.011

Source: Analysis of data from 26 districts in three provinces to calculate poverty and inequality measures using three methods: the Stata program used in this report, a SAS program developed by the World Bank with no adjustments for heteroskedasticity and location effects, and the same SAS program with adjustments for heteroskedasticity and location effects.

the Stata estimates and the SAS estimates with adjustment are the same, 23 percent and 17 percent of the SAS value, respectively.

Finally, Table A.4 shows the correlation coefficient ( $R^2$ ) between district-level estimates of rural poverty and inequality in the three provinces. The Stata poverty estimates are highly correlated with the corresponding SAS estimates, with  $R^2$  values of at least 0.995 in all cases. The Stata inequality estimates are less closely correlated with the

SAS estimates, with the  $R^2$  values varying between 0.68 and 0.94. The correlation coefficients, like the mean absolute differences, suggest that the Stata program measures the Gini index more accurately than the other two measures of inequality.

## Summary

In the first stage, the Stata program generates standard errors that are somewhat

**Table A.3 Comparison of poverty and inequality estimates using different methods**

Parameter	Stata (1)	SAS without adjustment (2)	SAS with adjustment (3)	Difference (1) – (2)	Difference (1) – (3)
Standard error of $P_0$ estimate					
Mean value	0.026	0.013	0.043	0.013	–0.017
Mean absolute value	0.013	0.017			
Estimate of poverty gap ( $P_1$ )					
Mean value	0.184	0.190	0.198	–0.006	–0.014
Mean absolute value	0.007	0.018			
Estimate of squared poverty gap ( $P_2$ )					
Mean value	0.090	0.094	0.098	–0.003	–0.008
Mean absolute value	0.005	0.009			
Estimate of Theil L index (GE(0))					
Mean value	0.112	0.103	0.091	0.009	0.021
Mean absolute value	0.009	0.021			
Estimate of Theil T index (GE(0))					
Mean value	0.109	0.104	0.093	0.005	0.016
Mean absolute value				0.005	0.016

Source: Analysis of data from 26 districts in three provinces to calculate poverty and inequality measures using three methods: the Stata program used in this report, an SAS program developed by the World Bank with no adjustments for heteroskedasticity and location effects, and the same SAS program with adjustments for heteroskedasticity and location effects.

larger than those of the SAS program, with or without adjustment for heteroskedasticity and location effects. This suggests that the Stata program does not exaggerate the statistical significance of the regression coefficients, at least in the case of estimating rural consumption in Vietnam.

In the second stage, the Stata estimates of  $P_0$  are, on average, about 3 percentage points different from those of the SAS program with adjustment, but there is little or no bias. However, the Stata program appears to underestimate the margin of error of these estimates by about 40 percent. The Stata

**Table A.4 Correlation ( $R^2$ ) between poverty and inequality estimates**

Parameter	Stata versus SAS	
	Without adjustment	With adjustment
Incidence of poverty ( $P_0$ )	1.000	0.995
Poverty gap ( $P_1$ )	0.997	0.996
Poverty gap squared ( $P_2$ )	0.997	0.996
Gini coefficient	0.937	0.783
Theil L [GE(0)]	0.829	0.682
Theil T [GE(1)]	0.914	0.744

Source: Analysis of data from 26 districts in three provinces to calculate poverty and inequality measures using three methods: the Stata program used in this report, an SAS program developed by the World Bank with no adjustments for heteroskedasticity and location effects, and the same SAS program with adjustments for heteroskedasticity and location effects.

estimates of  $P_1$  and  $P_2$  appear to be underestimated by about 7–8 percent compared to the corresponding SAS estimates with adjustment. The district-level poverty estimates produced by Stata and SAS (with adjustment) are highly correlated, with values of  $R^2$  of 0.995 and 0.996 for all three poverty indicators.

The Stata program overestimates inequality somewhat, though the degree varies across the three measures. The Stata estimates of the Gini coefficient are fairly accu-

rate, with an upward bias of 2 percent and an average difference of less than 5 percent compared to the SAS estimate with adjustment. The Stata estimates of Theil L and Theil T have larger upward biases: about 26 percent and 20 percent, respectively. The district-level inequality measures generated by Stata and SAS (with adjustments) are less strongly correlated than the poverty measures, with values of  $R^2$  between 0.68 and 0.78, depending on the index used.

## APPENDIX B

---

### Using GIS-Derived Variables for Statistical Analysis

#### General Considerations on GIS-Derived Variables

**T**his analysis of alternative determinants of poverty focuses on exploring linkages among district-level poverty incidence and a variety of agro-ecological variables by means of a regression analysis. Although the dependent variable is a result obtained from the Census and survey-based poverty mapping regression analysis, almost all of the independent explanatory variables used in this analysis were derived from GIS data layers. Some general aspects of the use of GIS-derived variables in statistical analysis are therefore considered here first.

An initial challenge in such an analysis is to establish an analytic link between the people and their local environment, that is, between the tabular socioeconomic data (the poverty incidence estimates at district level) and the GIS-based environmental data. One important facet of this challenge is that the dependent variable is of a different spatial data type and of a different form of spatial aggregation than most of the independent variables: socioeconomic variables typically exist in a spatially discrete representation format referring to administrative units or points, whereas environmental data are normally of a spatially continuous nature. This poses methodological challenges in spatial analysis applications, a problem generically known as the modifiable areal unit problem (MAUP). MAUP is endemic to all spatially aggregated data: the core issue is that alternative forms of spatial aggregation of data would lead to different results, or as Heywood (1998) put it: MAUP is “a problem arising from the imposition of artificial units of spatial reporting on continuous geographical phenomenon resulting in the generation of artificial spatial patterns.” Clearly, MAUP is an issue that needs to be kept in mind when socioeconomic and environmental data sets are to be combined, as attempted for the purpose of this analysis.

The successful generation of reliable estimates of district level poverty incidence allows a linkage analysis at the district level, which implies that agro-ecological variables need to be calculated at the district level. This, again, implies the need for some form of geographic aggregation of the spatially continuous data to spatially discrete record values at district level. The definition of the spatial extent (the areal unit) of the “district” now has considerable implications on the outcome of the aggregated data set and thus the analysis’ results. The agro-climatic variables to be linked to the socioeconomic data within each district area could be defined as, for instance and possibly most obviously, all observations within the administrative boundaries of a district, or, potentially more meaningfully, as those observations within a

certain perimeter of the populated areas of the district, or as the area within any other demarcation of the spatial extent of the district population's environment or area of direct socioeconomic activity.

Suppose there are two geographically similar districts with a lowland plain area and an adjacent mountainous hinterland. In one district 90 percent of the population is living as predominantly paddy farmers in the plain, whereas the vast majority of the population of the other district dwell as upland farmers in the highlands. Although the socioeconomic activities and natural resource uses, and hence the actual socioeconomic–environmental links, will most probably be very different in district A from those in district B, the populations of both districts will, as district aggregates, be statistically linked to a similar environment.

Clearly, the areal units for linking people to an environmental zone are largely arbitrary. In addition, the aggregated variables would have different values depending on the choice of the areal unit.

Although such areal unit aggregates are relatively easy to link with the socioeconomic data of the same areal unit, information on variations within the geographic unit, however spatially defined, is lost. For example, a district with a very rugged terrain with elevations ranging from 200 meters above sea level (mas) at valley bottoms to 2,700 mas at mountain peaks could have a district level mean elevation of 1,200 mas, similar to a district that is part of a plateau at 1,200 mas.

From the above, two things become obvious. First, alternative zoning of the areal units, possibly including some sort of spatial weighting scheme reflecting distributions of population and economic activities within the district (see, e.g., Epprecht and Müller 2003), promises an improvement of the analytic “people–environment” link. Second, the different aggregated variables need to represent the variations within the districts as well as possible. At the least, in addition

to means, additional variables representing the geographic variation of the same GIS variable (e.g., elevation) are essential.

The scope of this research study, however, only allowed for an adequate consideration of the latter. For this linkage analysis, therefore, mainly the mean, minimum, maximum, and standard deviation values for each district defined by its administrative boundaries were calculated for several of the spatially continuous GIS data, which attempt to reflect some of the variation of each variable within each district.

Nevertheless, because policy decisions are often made on the basis of results obtained from statistical analysis of at least partly spatially defined data (e.g., allocation of special funds to the poorest districts), the former needs to be kept in mind when attempting to draw conclusions from the results of the following analysis, and more attention needs to be paid to the problem in future analysis of spatial data.

The data sources, data processing, and the generated variables are described in more detail in the following sections.

## Data Sources

Data of different types and from a number of different sources were used to generate the variables used in the analysis. The digital elevation data used to generate related variables (see below) are based on *GTOPO30*, a global digital elevation model (DEM) with a horizontal grid spacing of 30 arc seconds (equivalent to approximately 1 kilometer), produced by the United States Geological Survey (USGS 2003). *GTOPO30* was derived from several raster and vector data sources of topographic information.

GIS data layers on transportation (roads and railroads) and river networks, as well as administrative boundaries, were obtained from the Centre for Remote Sensing and Geomatics (VTGEO) in Hanoi. Data on land cover, soil, and climate originated from the Ministry of Science and Technology (MOSTE).

## Data Processing and Data Quality

In the following, aspects of data quality and accuracy are discussed briefly, and the necessary data-processing tasks, as well as the data extraction methods, are outlined.

In regard to the sizes of the districts (the geographic unit to which the data were to be aggregated), an accuracy and therefore a target scale of the source maps of 1:250,000 appeared reasonable. However, for several data layers (soils and land cover in particular), only data based on source maps or equivalents (mainly satellite images) of scales no better than 1:500,000 or 1:1,000,000 were available. The USGS elevation data correspond to an accuracy of approximately 1:1,000,000.

The data obtained from the different sources described above were of varying quality, which made extensive processing tasks necessary. A general observation on GIS data in Vietnam is that basically all GIS data sets available on a national level are produced for an intended primary application in the field of cartography and mapping. Although the quality of the data is generally good enough for visualization on maps, the many topological errors often not visible on maps pose considerable problems when the data are to be used for GIS modeling and spatial analysis purposes. The greatest problems encountered were with digital line and polygon data (see below).

### Digital Elevation Data

Digital elevation data and derivatives such as slope, terrain roughness, and shaded relief (for illustrative mapping) were calculated using the above-mentioned GTOPO30 data as a basis. Although the data set, which comes with a declared vertical accuracy equivalent to an 18-meter root mean square error, was adequate for generating average elevation variables for each district, the generation of minimum, maximum, and standard deviation variables for elevation required some further processing of the data.

A process commonly referred to as filling of sinks was performed on the data set in order to eliminate unnatural sinks as a result of data errors. Still, district minimum and maximum elevation values extracted from those data were rather unreliable and were therefore not included in the variable database. GIS data layers describing slope as well as three different definitions of terrain roughness were calculated and modeled in a GIS environment. Slope is calculated by identifying the maximum rate of change in elevation value from each grid cell to its neighbors. Terrain roughness was modeled as the standard deviation of elevation values of all the neighboring cells in a given radius. In addition to these topographic derivatives, a shaded relief was produced for analytic and illustrative mapping purposes using the digital elevation data.

### Vector Line Data

The digital line data layers representing river and transportation networks were generally good enough for the calculation of proximity variables such as “average distance to a main road.” The many topological errors in those data sets, however, did not permit direct calculations of “distances along networks” variables such as “road distance to nearest town.” The line data were therefore converted into grid cell data, which permitted an approximate distance calculation using cost-distance modeling techniques. Transportation network density variables were calculated for each district by intersecting transportation data layers with the district boundary layer before calculating the total length of all roads per district as well as the total district area. Density variables were generated by dividing the former by the latter.

### Vector Polygon Data

Polygon data layers used for the analysis consisted mainly of layers on administrative boundaries, soil, and land cover maps. With the administrative boundaries, two challenges

were encountered. First, the administrative coding system had to be matched with the one used in the 1999 Housing and Population Census, whereas the relatively large number of changes in administrative divisions (splitting and merging of administrative units) added an extra challenge to this. Second, the many topological errors did not permit any areal calculations per administrative unit. This, however, is a necessary precondition for the creation of many of the district-level variables. Extensive spatial cleaning operations were therefore performed on the data sets. Although the administrative boundary data were kept in the polygon data structure, soil and land cover data sets were converted into grid data layers, which facilitated areal calculations and data extraction. In a GIS environment, “zonal statistics” (in this case statistics per district) were performed to generate variables on, for example, areas of each land cover class per district. Similar spatiostatistical analyses were performed using the elevation and derivative data layers with the administrative boundary polygons to calculate variables on, for example, mean, minimum, maximum, and standard deviation of elevation, terrain roughness, or slope per district.

### Vector Point Data

Digital point data used in the analysis included point data layers on location of administrative centers and on the location of climatic measurement stations, including attribute tables describing monthly averages of rainfall, temperature, humidity, and sunshine duration.

*Climatic Data.* In order to be able to generate climatic indicator variables for each district, the values of the variables available for the climate measurement stations should be available for each of the 614 districts rather than only for the 161 measurement stations. Because such information is not available, spatial interpolation techniques were applied to calculate climate variable surfaces for the whole of Vietnam, from

which district aggregated values can then be calculated. In order to produce accurate climate variable surfaces, sophisticated modeling techniques would have to be applied, taking into account factors such as elevation, landform, and land cover, which have a direct influence on local climatic conditions. However, such sophisticated climatic modeling procedures clearly would have gone far beyond the scope of this study. Therefore, relatively straightforward interpolation techniques were applied, using Kriging techniques. Kriging is an advanced geostatistical procedure that generates an estimated surface from a scattered set of points with  $z$  values. Unlike other interpolation methods, Kriging involves an interactive investigation of the spatial behavior of the phenomenon represented by the  $z$  values before selecting the best estimation method for generating the output surface.

*Towns and Cities.* Those data sets were used in the form of points as well as polygon data. Although the point data layer on the location of the administrative centers per administrative unit (province and district) required updating predominantly for the places where geographic administrative changes occurred, the categorization of the towns and cities by size required additional inputs. In order to determinate the size (number of people) and the areal extent of the town and cities (many towns, and cities in particular, extend across districts), information from the 1999 Housing and Population Census was integrated into the GIS database. By combining information on urban areas and population numbers available in the Census with the respective administrative units in the GIS database, delineation of the urban areas and calculation of their population was possible. For calculations of distances to towns or cities, the distance to the nearest urban perimeter was taken as the respective reference. Elevation values of towns were obtained by overlaying the GIS point data layers of the town locations over the gridded digital elevation model and ex-

tracted through an assignment of data by location process.

### Description of Variables

In total, some 430 individual variables were derived from GIS data layers. In addition, another 22 variables on markets and population numbers originating from published results of the 1999 Vietnam Market Networks survey and from the 1999 Population and Housing Census, respectively, were included in the database. The following list provides an overview of the variables, grouped into six broad categories:

#### Location

- UTM48 projected XY coordinates of district towns
- UTM48 projected XY coordinates of district centroids

#### Socioeconomic Variables

- Number of communes
- Population of districts (total, male, female)
- Population of the district capitals
- Number of markets
- Markets per commune (district average)
- Market revenues and payments of planned markets to the state

#### Land Cover

- Total area of the district
- Arable land area
- Natural forest area
- Planted forest area
- Bare and rocky land area

#### Relief

- Percentage of total area by elevation range (0–250 meters, 250–500 meters, 500–1,000 meters, 1,000–1,500 meters, over 1,500 meters)
- District elevation values (mean, median, standard deviation)
- Elevation of district town
- Roughness as standard deviation of cell values in a radius of 5, 12, and

25 kilometers (minimum, maximum, range, mean, standard deviation)

- Percentage of total district area by slope class (0–4 percent, 4–8 percent, 8–15 percent, 15–30 percent, over 30 percent)
- Percentage of total district area by soil type (32 soil type classes)

#### Transportation and Accessibility

- Length of roads by type (total, main roads, minor roads, tracks)
- Navigable rivers total length
- Distance to main road (maximum, mean, standard deviation)
- Distance to a road (maximum, mean, standard deviation)
- Distance from district town to closest province town
- Average distance to province town (minimum, maximum, mean, standard deviation)
- Average distance to district town (maximum, mean, standard deviation)
- Distance from district town to town with over 10,000, 50,000, 100,000, 250,000, 1 million inhabitants
- Distance to town with over 10,000, 50,000, 100,000, 250,000, 1 million inhabitants (minimum, maximum, mean, standard deviation)
- Approximate distance along best available roads from district town to nearest town with over 10,000, 50,000, 100,000, 250,000, 1 million inhabitants
- Approximate time along best available roads from district town to nearest town with over 10,000, 50,000, 100,000, 250,000, 1 million inhabitants

#### Climate

- Monthly average precipitation (minimum, maximum, range, mean, standard deviation)
- Monthly average temperature (minimum, maximum, range, mean, standard deviation)



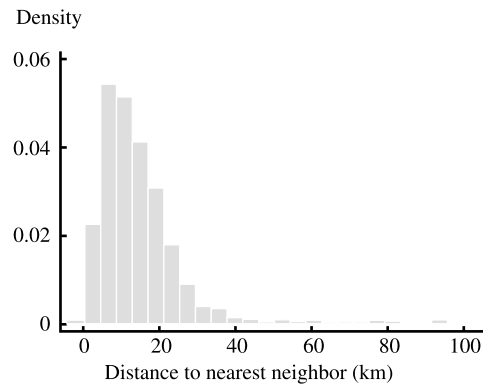
- Monthly average sunshine duration (minimum, maximum, range, mean, standard deviation)
- Monthly average humidity (minimum, maximum, range, mean, standard deviation)

### Spatial Weighting Matrix

In the analysis of spatially derived data, a phenomenon generally referred to as spatial autocorrelation needs to be considered. Spatial autocorrelation is a spatial dependence problem based on the similarity of nearby observations (i.e., the propensity for data to be similar to surrounding data values), which contradicts the common statistical assumption of independence of observations (see further details below). To correct for such distorting effects in the regression analysis, which would lead to biased results, a common practice is to construct a spatial weighting matrix. With a spatial weighting matrix, an attempt is made to quantify the often subjective concept of proximity (there is no standard on how to conceptualize and quantify “nearness” or “relatedness”). Two common types of spatial weighting matrices are generally distinguishable: (1) discrete contiguity matrixes with values 1 and 0 depending on whether polygons are contiguous and (2) continuous spatial weighting matrixes where observations are weighted by some distance decay function.

For the purpose of this analysis, a weighting scheme based on inverse distance was chosen, in which the distance measurements were taken from the district centroid points. In order to specify the distance after which threshold weights become 0, a distance band within which location pairs must be considered “neighbors” (i.e., spatially contiguous) must be defined. The choice of the distance

**Figure B.1 Distance to nearest neighboring district measured from district centroids**



band has direct implications on the degree of weighting decay by distance. The distance band needs to be defined in such a way that each observation has at least one “neighbor.” Figure B.1 shows the histogram for straight-line distances from each district centroid to its nearest neighbor.<sup>29</sup> Clearly, there are a small number of observations with exceptionally large distances to the closest neighbor. Not surprisingly, the top three districts in terms of distance to the nearest neighbor are island districts off the mainland coast. Dropping those three island districts (Bach Long Vi District, Hai Phong Province; Phu Qui District, Binh Thuan Province; Con Dao District, Ba Ria–Vung Tau Province) from the analysis allowed for a maximum distance band specification of 75 kilometers so that each district has at least two “neighbors.” A too-large distance band specification would result in a too-weak correction of local spatial autocorrelation effects, whereas a too-low upper limit of the distance band would create a spatial weighting matrix in which some observations would have no “neighbors.”

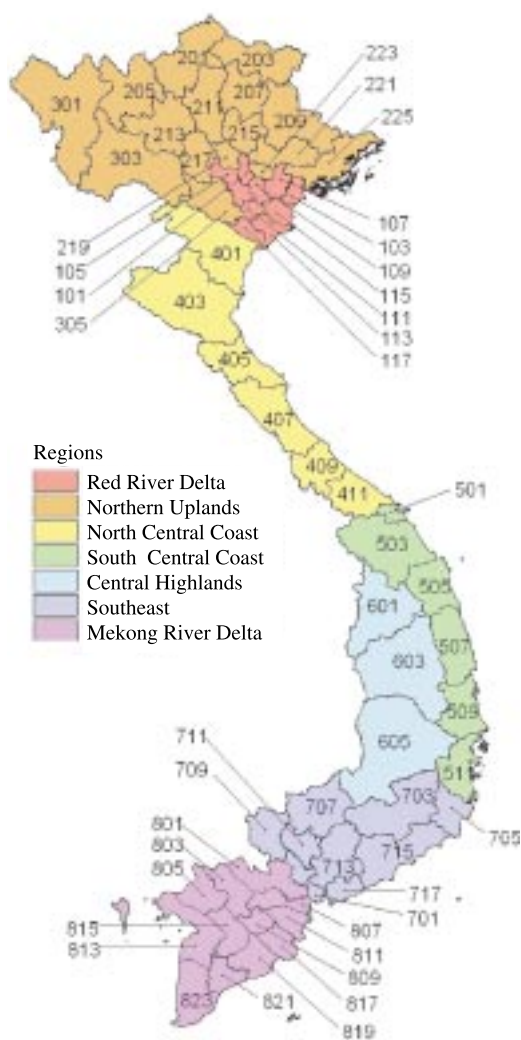
<sup>29</sup>The graphic shows the nearest-neighbor distances of all the 614 districts included in the poverty-mapping analysis. The two archipelago districts Hong Sa and Truong Sa, both well over 300 km away from the closest mainland district, were thus excluded from the analysis.

## References

- Anselin, L., 1988. *Spatial econometrics: Methods and models*. Dordrecht: Kluwer.
- Baker, J., and Grosh, M. 1994. Poverty reduction through geographic targeting: How well does it work? *World Development* 22 (7): 983–995.
- Baulch, B. 2002. Poverty monitoring and targeting using ROC curve: Examples from Vietnam. IDS Working Paper 161. Brighton: Institute of Development Studies.
- Bigman, D., and H. Fofack. 2000. *Geographic targeting for poverty alleviation: Methodology and applications*. Washington, D.C.: World Bank Regional and Sectoral Studies.
- Brunsdon, C., A. S. Fotheringham, and M. E. Charlton. 1996. Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographic Analysis* 28 (4): 281–298.
- Conway, T. 2001. *Using government data to target activities to poor communes and monitor poverty reduction: A review of options for the Cao Bang-Bac Kan Rural Development Project*. Hanoi: Commission of the European Communities.
- Elbers, C., J. Lanjouw, and P. Lanjouw. 2003. Micro-level estimation of poverty and inequality. *Econometrica* 71 (1): 355–364.
- . 2004. *Imputed welfare estimates in regression analysis*. Working Paper Series No. 3294. Washington, D.C.: World Bank.
- Epprecht, M., and D. Müller. 2003. Linking people and the landscape: GIS and spatial analytical techniques for agricultural economists, presentation at the International Conference of International Association of Agricultural Economists, Durban, 2003.
- Fotheringham A. S., M. Brunsdon, and M. Charlton. 2002. *Geographically weighted regression: The analysis of spatially varying relationships*. Chichester: Wiley.
- GSO (General Statistics Office). 2000. *Vietnam living standards survey 1997–1998*. Hanoi: Statistical Publishing House.
- Henninger, N., and M. Snel. 2002. *Where are the poor? Experiences with the development and use of poverty maps*. Washington, D.C.: World Resources Institute; Arendal, Norway: UNEP-GRID/Arendal.
- 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 (1): 147–165.
- Heywood, I. 1998. *Introduction to geographical information systems*. New York: Addison-Wesley Longman.
- Kanbur, R. 2002. Notes of the policy significance of inequality decompositions. Cornell University, Ithaca, N.Y., U.S.A. Mimeo.
- Minot, N. 1998. *Generating disaggregated poverty maps: an application to Viet Nam*. Markets and Structural Studies Division, Discussion Paper No. 25. Washington, D.C.: International Food Policy Research Institute.
- . 2000. Generating disaggregated poverty maps: An application to Vietnam. *World Development* 28 (2): 319–331.

- Minot, N., and B. Baulch. 2002a. *The spatial distribution of poverty in Vietnam and the potential for targeting*. Markets and Structural Studies Division Discussion Paper No. 42. Washington, D.C.: International Food Policy Research Institute.
- . 2002b. *Poverty mapping with aggregate census data: What is the loss in precision?* Markets and Structural Studies Division Discussion Paper No. 49. Washington, D.C.: International Food Policy Research Institute.
- Poverty Working Group. 1999. *Vietnam development report: Attacking poverty*. Joint report of the Government of Vietnam–Donor-NGO Poverty Working Group presented to the Consultative Group Meeting for Vietnam.
- StataCorp. 2001. *Stata reference manual*, release 7, Volume 4, svymean. College Station, Tex., U.S.A.: Stata Press.
- Statistics South Africa and the World Bank. 2000. Is census income an adequate measure of household welfare: Combining census and survey data to construct a poverty map of South Africa. Washington, D.C. Mimeo.
- USGS (United States Geological Survey). 2003. GTOPO30. <http://edcdaac.usgs.gov/gtopo30/gtopo30.html>, accessed 2003.
- World Bank. 2000. *Panama poverty assessment: Priorities and strategies for poverty reduction*. Washington, D.C.: World Bank Country Study.

**Figure 1.1 Regions and provinces of Vietnam**



**Provinces**

101 Ha Noi	225 Quang Ninh	703 Lam Dong
103 Hai Phong	301 Lai Chau	705 Ninh Thuan
105 Ha Tay	303 Son La	707 Binh Phuoc
107 Hai Duong	305 Hoa Binh	709 Tay Ninh
109 Hung Yen	401 Thanh Hoa	711 Binh Duong
111 Ha Nam	403 Nghe An	713 Dong Nai
113 Nam Ding	405 Ha Tinh	715 Binh Thuan
115 Thai Binh	407 Quang Binh	717 Ba Ria-Vung Tau
117 Ninh Binh	409 Quang Tri	801 Long An
201 Ha Giang	411 Thua Thein-Hue	803 Dong Thap
203 Cao Bang	501 Da Nang	805 An Giang
205 Lao Cai	503 Quang Nam	807 Tien Giang
207 Bac Kan	505 Quang Ngai	809 Vinh Long
209 Lang Son	507 Binh Dinh	811 Ben Tre
211 Tuyen Quang	509 Phu Yen	813 Kien Giang
213 Yen Bai	511 Khanh Hoa	815 Can Tho
215 Thai Nguyen	601 Kon Tum	817 Tra Vinh
217 Phu Tho	603 Gia Lai	819 Soc Trang
219 Vinh Phuc	605 Dak Lak	821 Bac Lieu
221 Bac Giang	701 Ho Chi Minh	823 Ca Mau
223 Bac Ninh		

**Figure 3.1 Map of the incidence of poverty ( $P_0$ ) for each province**

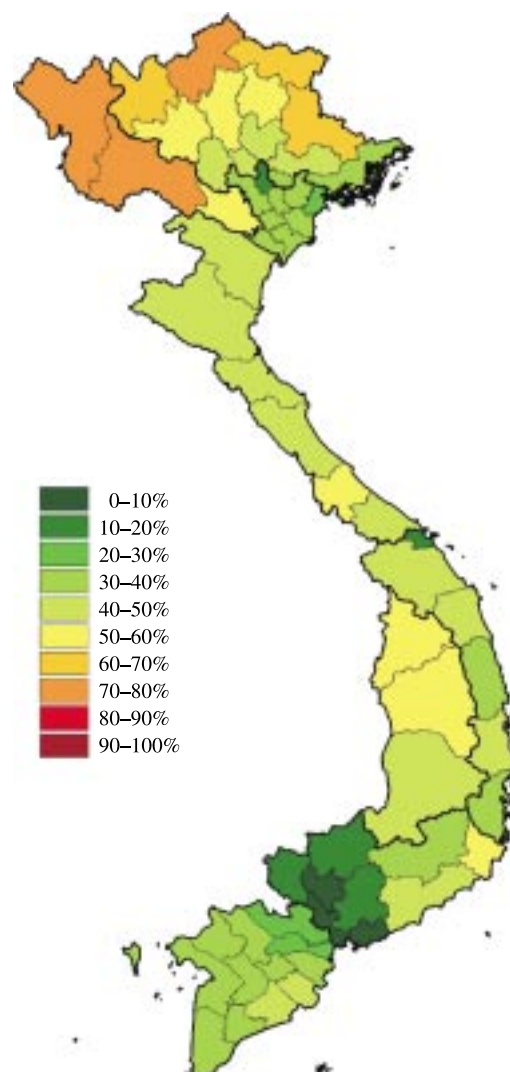


Figure 3.3 Map of the incidence of poverty ( $P_0$ ) for each district

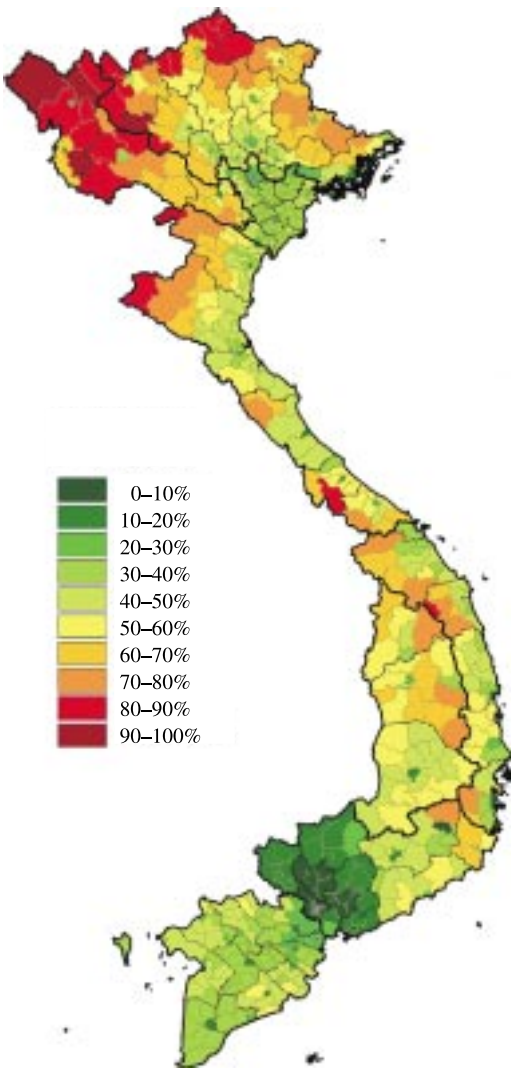
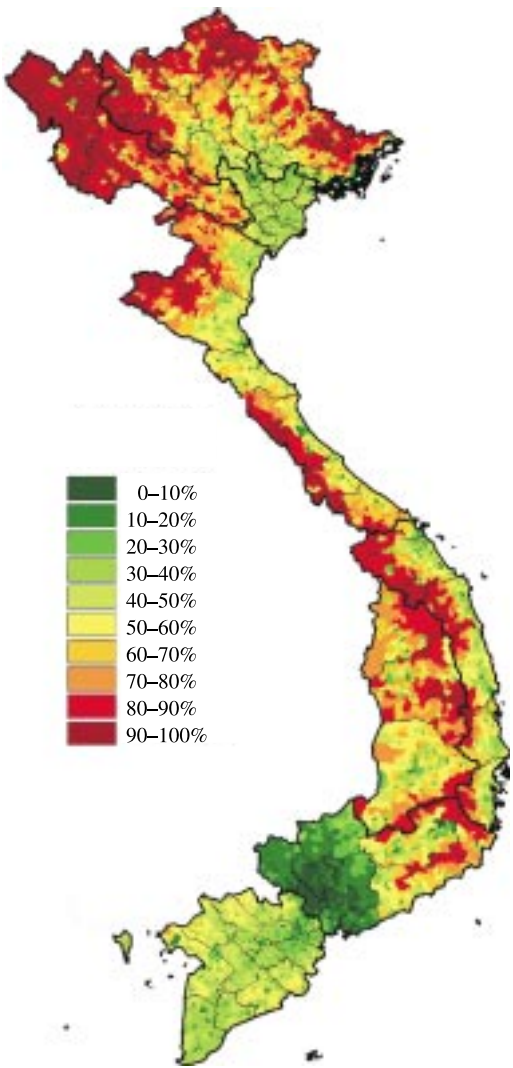
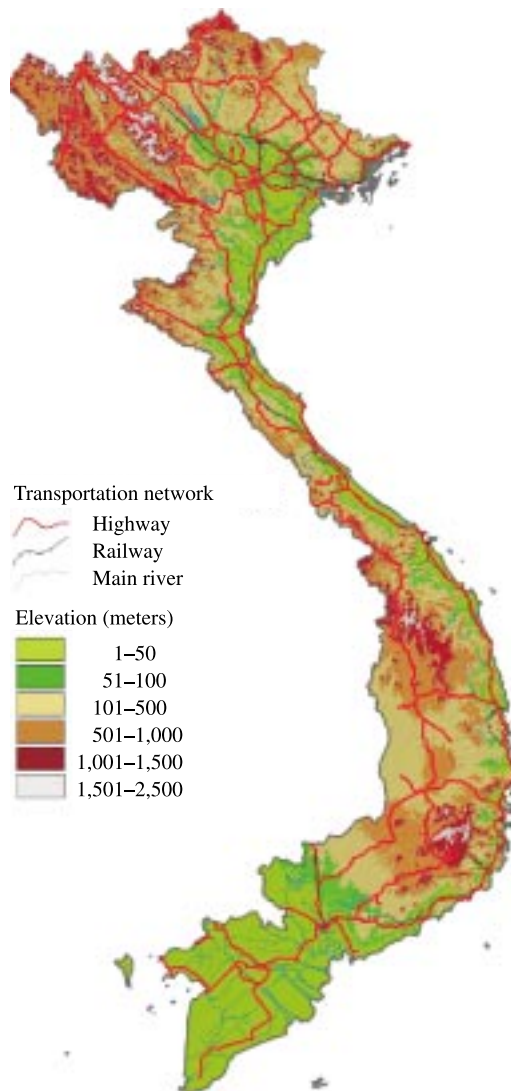


Figure 3.6 Map of the incidence of poverty ( $P_0$ ) for each commune



**Figure 3.7 Map of elevation and transportation infrastructure**



**Figure 3.9 Map of the density of poverty**

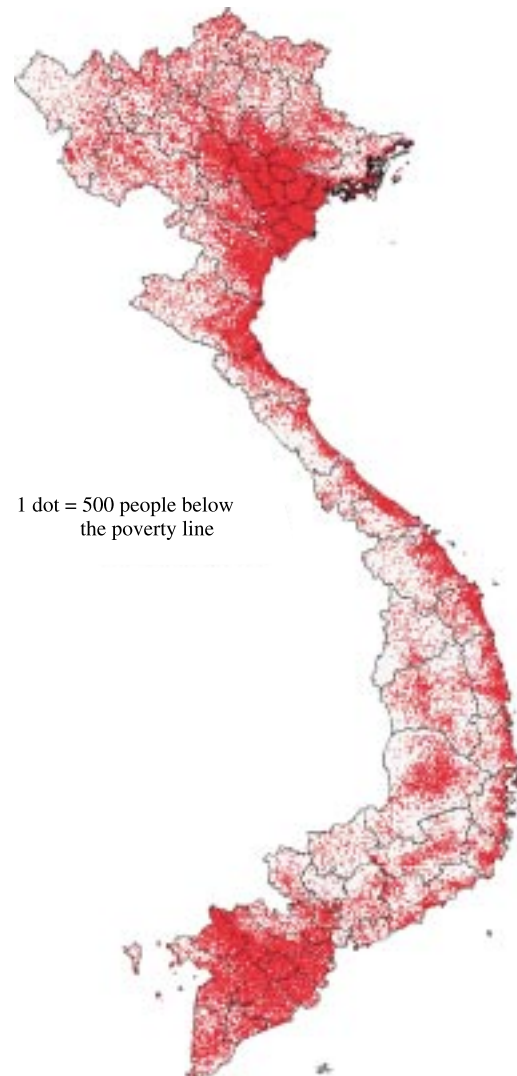
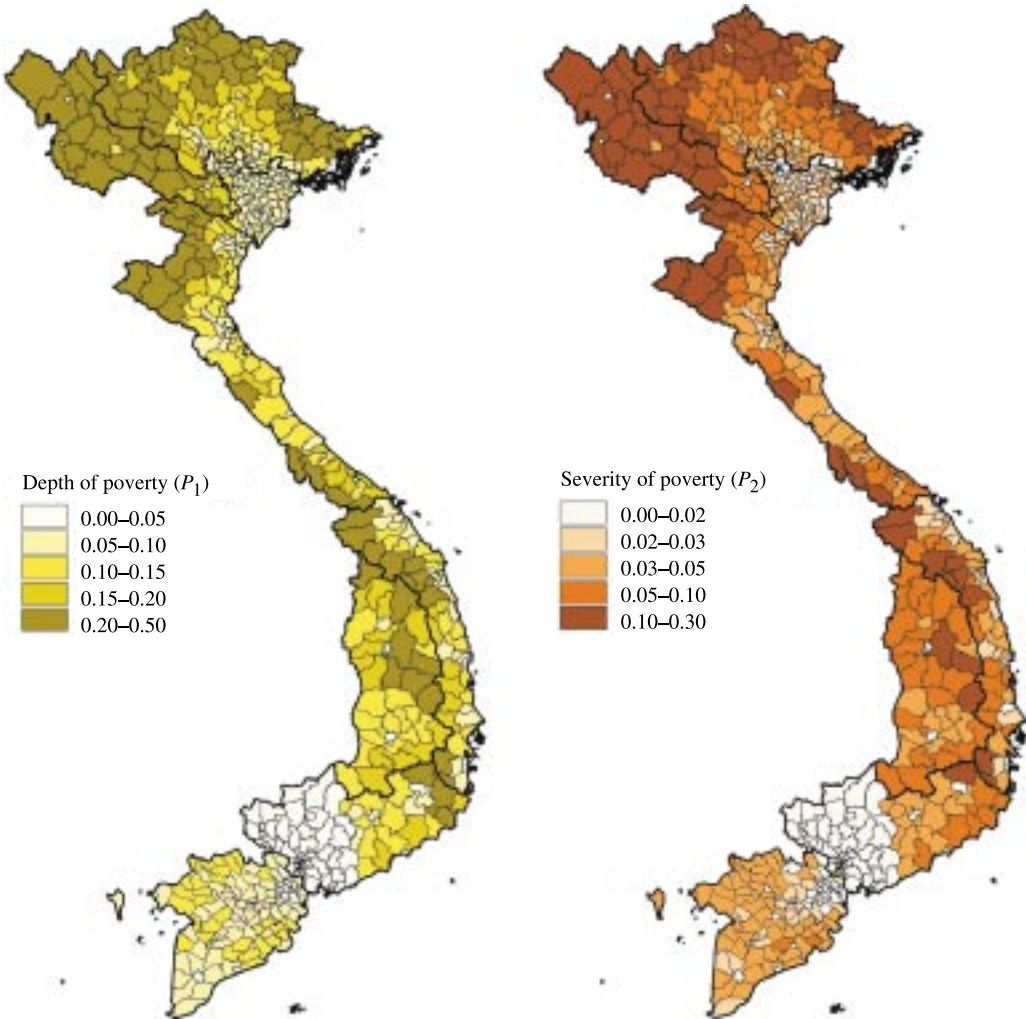
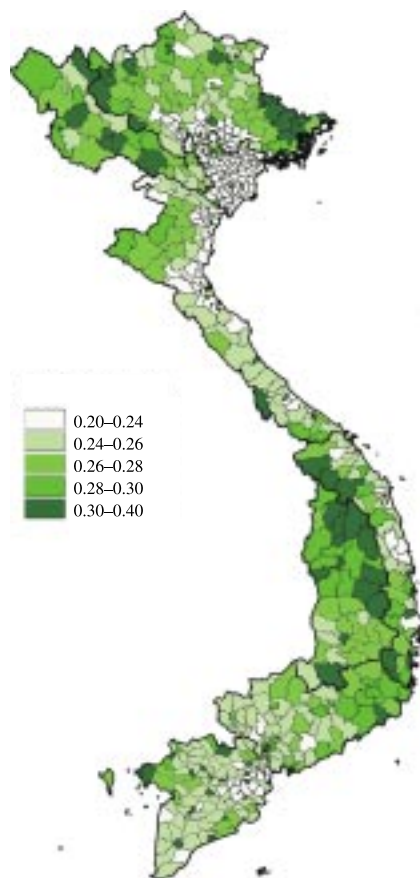


Figure 3.10 Maps of the depth of poverty ( $P_1$ ) and severity of poverty ( $P_2$ ) for each district

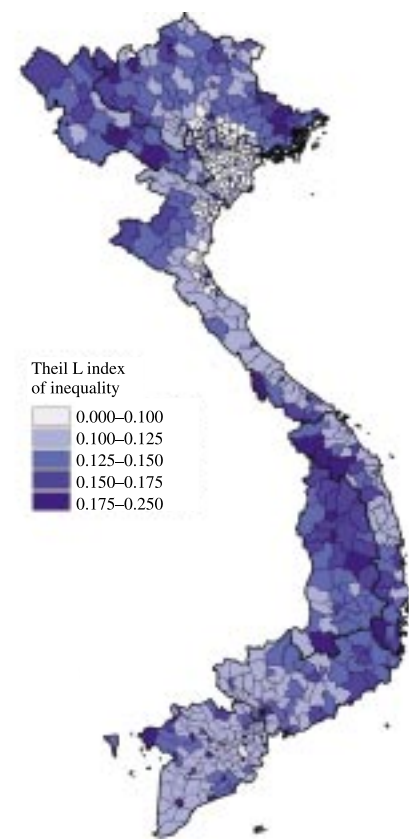
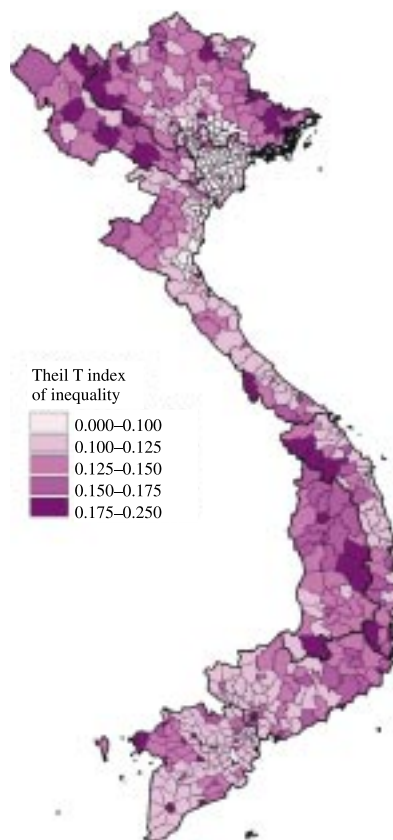




**Figure 3.12 Map of inequality as measured by the Gini coefficient**



**Figure 3.13 Maps of inequality as measured by the Theil L and Theil T indexes**





**Figure 5.1 Maps of the spatial distribution of the independent variables**

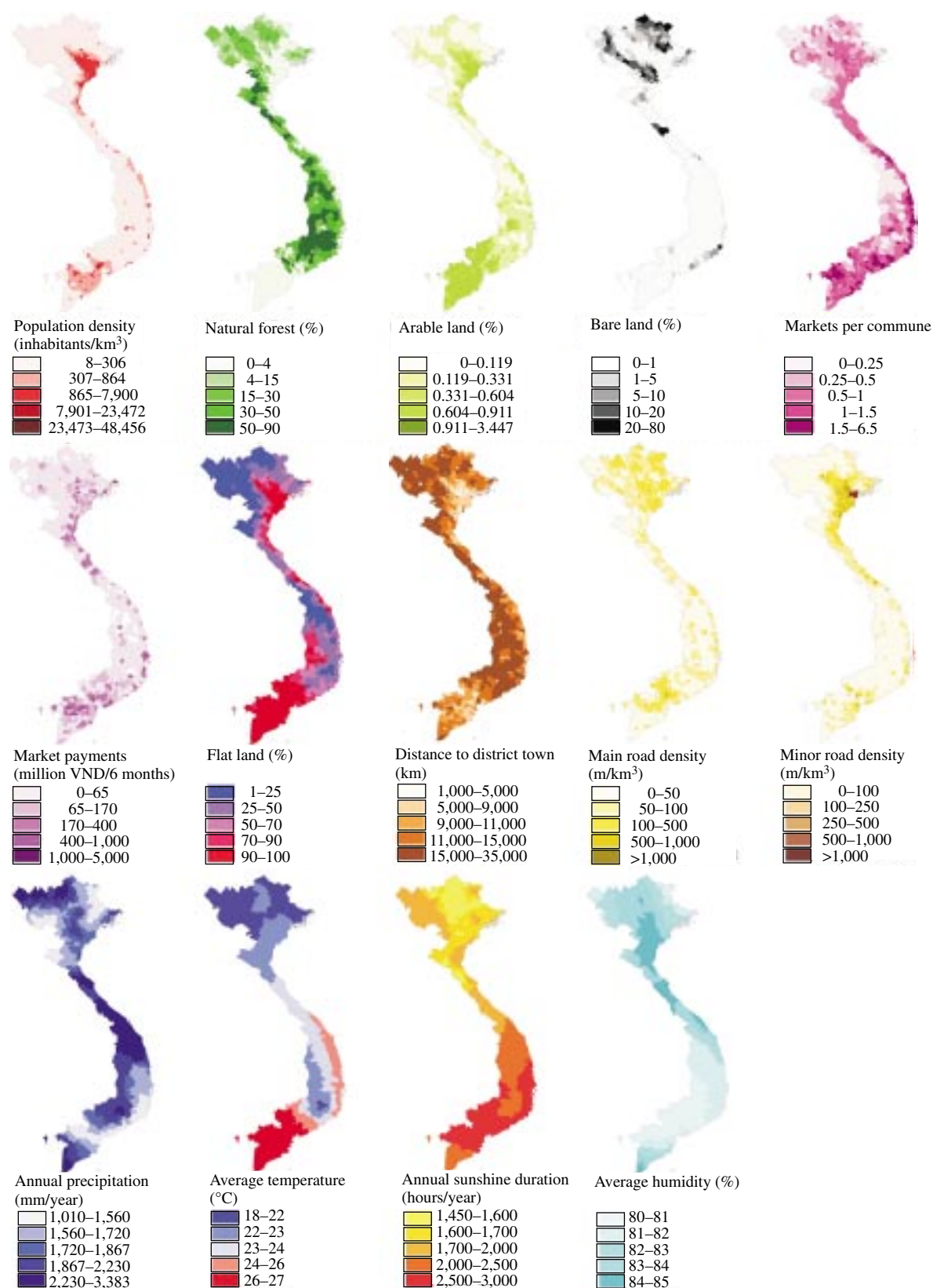


Figure 5.3 Map of the spatial distribution of local  $R^2$

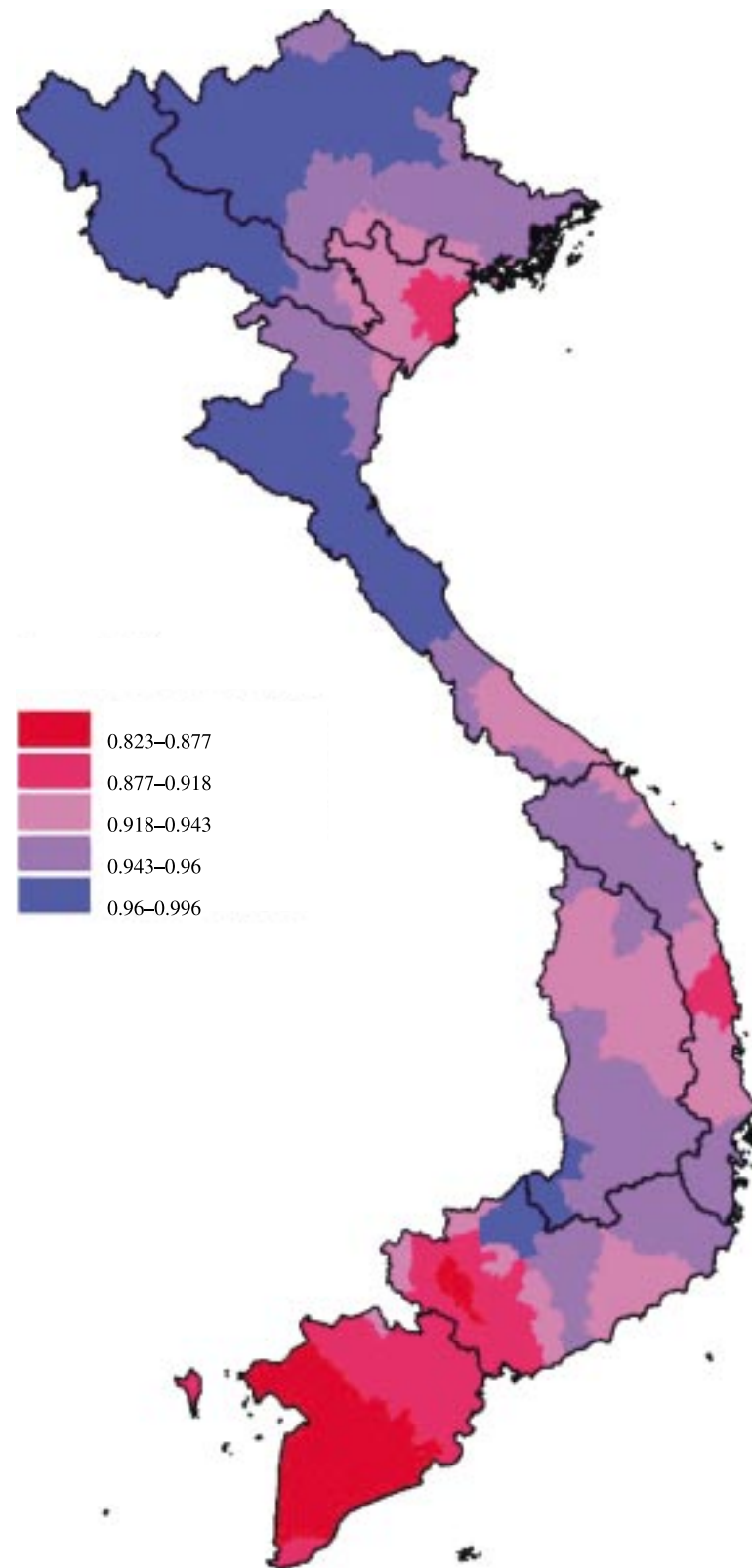


Figure 5.4 Maps of the spatial distribution of the local coefficients of the independent variables

