Metodološki zvezki, Vol. 1, No. 1, 2004, 225-234

Autologistic Regression Model for Poverty Mapping and Analysis

Alessandra Petrucci, Nicola Salvati, and Chiara Seghieri[1]

Abstract

Poverty mapping in developing countries has become an increasingly important tool in the search for ways to improve living standards in an economically and environmentally sustainable manner. Although the classical econometric methods provide information on the geographic distribution of poverty, they do not take into account the spatial dependence of the data and generally they do not consider any environmental information. Methods which use spatial analysis tools are required to explore such spatial dimensions of poverty and its linkages with the environmental conditions. This study applies a spatial analysis to determine those variables that affect household poverty and to estimate the number of poor people in the target areas.

1 Introduction

Poverty maps are important tools that provide information on the spatial distribution of poverty within a country. They are used to affect various kinds of decisions, ranging from poverty alleviation programmes to emergency response and food aid.

However, the use of poverty maps alone does not furnish an estimate of the causal linkage between poverty and the variables influencing it; such maps furnish only "visual" advice. For this reason, researchers usually look for the possible existence of empirical relationships between poverty and socio-economic indicators. They make use of statistical methods such as the econometric model that combines census and survey data as applied in South Africa and Ecuador (Hentschel et al., 2000).

Generally poverty map studies don't take account of the geographical components (location) and environmental data that may have an important impact on research results. Environmental degradation contributes to poverty through worsened health and by constraining the productivity of those resources on which the poor rely. Moreover, poverty restricts the poor to acting in ways that harm the environment. Poverty is often concentrated in environmentally fragile ecological zones where communities face and contribute to different kinds of environmental degradation. In addition, demographic factors can be involved in complex ways (high population growth rates are associated with poverty) and exacerbate problems of environmental degradation directly.

[1] University of Florence - Department of Statistics "G. Parenti" - Viale Morgagni, 59 - 50134 Florence - Italy; alex@ds.unifi.it salvati@ds.unifi.it seghieri@ds.unifi.it

The other hand, in the social sciences, spatial contiguity in social and economic variables is a consequence of the instincts of individuals and of the patterns of behaviour and economic constraints that taken together help bind social space into recognizable structures. In a village or urban community, many of the households may have similar sources of income, and all households are affected by the same agroclimatic and geographic conditions. They also have other circumstances in common including road conditions, availability of public facilities for services such as health, water supply and education. Hence, it is reasonable to suppose that households living in the same area tend to act in similar ways and to influence one another.

Therefore, methods which use spatial analysis tools are required to explore such spatial dimensions of poverty and its linkages with the environmental conditions. This study investigates an approach based on the spatial regression model, for mapping poverty in Ecuador.

2 Generalized spatial linear models

This study applies a spatial analysis to determine those variables that affect household poverty and to estimate the number of poor people in the target areas. This type of analysis is based on the assumption that measured geographic variables often exhibit properties of spatial dependency (the tendency of the same variables measured in locations in close proximity to be related) and spatial heterogeneity (non-stationarity of most geographic processes, meaning that global parameters do not well reflect processes occurring at a particular location). While traditional statistical techniques have treated these two last features as nuisances, spatial statistics considers them explicitly.

As a special case, generalized spatial linear models include spatial linear regression and analysis of variance models, spatial logit and probit models for binary responses, loglinear models and multinomial response models for counts.

Let c_i denote the level of consumption per household, z denote the poverty line, and $s_i = \frac{c_i}{z}$ be the normalized welfare indicator per household. The household poverty indicator is determined by the normalized welfare function as follows:

$$y_i = 1 \quad ln(s_i) < 0$$

$$y_i = 0 \quad ln(s_i) \geq 0$$

The households are observed in n sites that form a subset S of the space. Each point (household) i has a binary response y_i and a vector $k \times 1$ of covariates $\boldsymbol{x}_i$. The responses constitute a map $Y = (y_i)_{i=1}^{n}$.

The regression model is called autologistic and states the conditional probability p_i that y_i is equal to 1, given all other site values y_j $(j \neq i)$:

$$p_i = Pr(y_i = 1|y_j, j \neq i) = Pr(y_i = 1|y_j, j \in N(i)) = \Phi(\beta_0 + \boldsymbol{\beta}^H \boldsymbol{x}_i^H + \boldsymbol{\beta}^C \boldsymbol{x}_i^C + \gamma y_i^*) \tag{2.1}$$

where $N(i)$ is the neighbour set of site i according to a neighbourhood structure, $\boldsymbol{\beta}^H$ and $\boldsymbol{\beta}^C$ are the vectors of regression coefficients and y_i^* is the sum of the values of the dependent variable of the neighbours of the site i, that is:

$$y_i^* = \sum_{j=1}^{n} y_j I(i \cong j) = \sum_{j:i \cong j} y_j \tag{2.2}$$

where $i \cong j$ denotes that the households i and j are neighbours.

This kind of model takes into account the spatial distribution of the welfare indicator, incorporating the neighbourhood structure in the model as another parameter to estimate.

In the model, $\boldsymbol{X}^H$ is the vector of explanatory variables that describe the household characteristics, $\boldsymbol{X}^C$ is the vector of explanatory variables describing the characteristics of the area in which the households reside, and Φ is a cumulative distribution function that is standard normal in the case of probit regression.

For a given poverty line and a given set of observation on $\boldsymbol{X}^H$ and $\boldsymbol{X}^C$, the estimates of $\boldsymbol{\beta}^H$, $\boldsymbol{\beta}^C$ and γ can be obtained by the maximum pseudo-likelihood method. Besag (1975) has demonstrated that the pseudo-likelihood method produces consistent parameter estimates under regular conditions.

Given the above generalized linear model, a maximum pseudo-likelihood estimator (MPE) for the unknown parameter vector $\boldsymbol{\Theta} = \{\beta_0, \boldsymbol{\beta}^H, \boldsymbol{\beta}^C, \gamma\}$ will be defined as the vector $\hat{\boldsymbol{\Theta}}$ that maximizes the pseudo-likelihood function:

$$\prod_{i=1}^{n} Pr(y_i = 1 | y_j, j \neq i) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{y_i} \tag{2.3}$$

As a result, the function in Equation 2.3 is not a full likelihood. An analytical form of the full likelihood is intractable for this problem because there is generally an unknown normalizing function (Besag, 1974). Note that the logit expressions in equation (2.3) are not independent across households because each household's variable y_i is related to the y_i variables of all the other households. Consequently, from the standpoint of the maximum likelihood estimation theory, we should not multiply the n logit likelihoods generated by equation (2.3) together to compute the overall likelihood function. Nevertheless, it can be shown that maximizing the function obtained by multiplying together the logit likelihoods represented by equation (2.3) yields consistent estimates of model parameters. This procedure, known as maximum pseudo-likelihood estimation (MPLE) (Cressie, 1993), provides consistent estimates of model parameters.

For the autologistic model, this approach is computationally simple since it amounts to using standard logit software to estimate the model parameters - ignoring the fact that the response variables are actually interdependent.

Therefore, the pseudo-likelihood estimation procedure proposed is an intuitively plausible method that avoids the technical difficulties of the full maximum likelihood approach. A drawback of the method is that its sampling properties have not been studied as extensively as those of the full maximum likelihood estimators.

Besag (1977) discusses the consistency and efficiency of pseudo-likelihood estimation for simple spatial Gaussian schemes. Strauss and Ikeda (1990) have shown that, for a logit model, maximization of Equation 2.3 is equivalent to a maximum likelihood fit for a logit regression model with independent observations y_i. Consequently, estimates can be obtained by using an iteratively reweighted least squares procedure. Therefore, any standard logistic regression routine can be used to obtain MPEs of the parameters. However, the standard errors of the estimated parameters calculated by the standard programs are not directly applicable because they are based on the assumption of independence of the observations.

The next step is the estimation of the incidence of poverty in all counties. These estimates are made on the basis of the relationship between the area characteristics and the probability that households residing in these areas are poor. The probability that households in a given county are poor is estimated only on the basis of the area characteristics:

$$\boldsymbol{\pi}_C = \Phi(\bar{\boldsymbol{X}}^C \boldsymbol{\beta}^H + \boldsymbol{X}^C \boldsymbol{\beta}^C) \tag{2.4}$$

where $\bar{\boldsymbol{X}}^C$ is a vector of variables describing the household characteristics calculated at area level, $\boldsymbol{\beta}^H$ and $\boldsymbol{\beta}^C$ are the coefficients from Equation 2.1 and $\boldsymbol{\pi}_C$ is the probability that a household drawn from a certain county is poor. The parameter estimates from the regression are applied to the census data in order to obtain an imputed value for $\boldsymbol{\pi}_C$, the percentage of poor households in a county. In this way, the poor households in all the counties are estimated. Finally, using the information on household size, the probability of a household being poor can be extended to the probability of an individual being poor.

3 The data

The first source of data considered for the purposes of this study is the Encuesta Condiciones de Vida (ECV) database. This database stems from a large and comprehensive survey conducted on Ecuador in 1995 that forms part of the World Bank's Living Standard Measurement Surveys (LSMS) project that started in 1980. The survey was administered to a sample of about 5 800 nationally representative households (this study makes use of a reduced sample of 5 630 households because data were missing for some variables). It collected data on all dimensions of household well-being and socio-economic characteristics including highly disaggregated data on household consumption expenditures. The survey design incorporated both clustering and stratification on the basis of the country's three main agroclimatic zones and rural-urban breakdown. It also oversampled Ecuador's two main cities: Quito and Guayaquil.

According to the ECV sampling design, the sample employed in this study is representative of the main agroclimatic zones. The sample size is too small to allow an estimation of the incidence of poverty at the level of provinces, counties and parroquias (municipalities). On the basis of this survey, if traditional mapping methods of spatial distribution of poverty were applied, the only working level should be the main regions in Ecuador (first level). However, by using data from another source, the INFOPLAN database, and aggregating the two databases at the common level of the county, it was possible to map the spatial distribution of poverty at county level.

INFOPLAN is an atlas that collects about 104 variables from the "Census of population and households" (INEC) conducted in Ecuador in 1990. It provides a wide variety of information on the demographic, socio-economic and geographical characteristics of the areas, and the data are available for many geographical area levels (from regions to parroquias). However, it does not contain income or consumption expenditure information for each household. Although the 1995 ECV data were collected five years after the census, the 1990–95 period was one of relatively slow growth and low inflation in Ecuador, so it is reasonable to assume that there was relatively little change.

Furthermore, the household living standards in the available counties are not georeferenced: the location of the respondent household was identified only by the county of residence and the type (rural or urban) of living area. This problem was overcome by locating each family randomly (assuming a uniform distribution) in the county but taking into account the type of living area. In order to understand the relationship between poverty and environment, the study also considered some environmental variables at the county level, e.g. cereal production, amount of arable land, and the distance of the households from the main roads (data provided by FAO/SDRN GIS).

Finally, all the data from the three sources of information (ECV, INFOPLAN and FAO) were arranged in a Geographic Information Systems (GIS) for managing the spatial dimension. The FAO and INFOPLAN data were already organized as GIS data and could be overlaid by merging the information contained in the layers. From the ECV data, a point coverage was created taking into account the geographical constraints.

4 Empirical results

From the methodological point of view, the spatial analysis is based on three steps:

- Step I. The spatial estimation of the impact of location characteristics of the areas in which the households reside is used to calculate the probability that these households are poor. The household data from the ECV and the community data from INFOPLAN are employed in order to determine the variables that best explain household consumption and poverty.

- Step II. Basic exploratory data analysis (EDA) techniques are applied and the spatial neighbourhood structure is defined in order to test the presence of spatial autocorrelation among the observed values.

- Step III. The incidence of poverty in all the target areas (counties) in the country (Ecuador) is estimated on the basis of their location-specific characteristics and on the relationship estimated in Step I.

Table 1 reports the estimated coefficients of the spatial probit regression model. It estimates the probability that the household is poor as a function of various households' characteristics, various characteristics of the area in which the household resides and of a component ($\mathbf{y}^*$) standing for the spatial dimension. Table 1 reports

Table 1: Coefficient estimates, standard error, z-value of the autologistic model.

Coefficients	*Estimate*	*Std.Error*	*z value*	$Pr(>\vert z\vert)$
% adults illiterate in hh	$2.69E-02$	$1.02E-01$	0.263	0.79
% persons with diploma	$-2.12E+00$	$2.49E-01$	−8.519	$2.00E-16$
Adequate home	$-7.62E-02$	$6.33E-02$	−1.204	0.22
Home with drinking-water	$-6.85E-03$	$5.01E-02$	−0.137	0.89
Home with adequate toilet	$-1.07E-01$	$6.20E-02$	−1.727	0.08
Home with adequate wall	$-1.83E-01$	$4.68E-02$	−3.91	$9.24E-05$
H. with public electricity n.	$8.40E-02$	$7.28E-02$	1.154	0.24
Waste collection by truck	$-3.13E-01$	$5.24E-02$	−5.977	$2.28E-09$
Persons per room	$3.75E-01$	$1.69E-02$	22.142	$2.00E-16$
Population	$7.98E-06$	$2.99E-06$	2.67	0.0075
Mortality rate (per 1000)	$5.81E-03$	$2.48E-03$	2.347	0.0189
Number of babies	$-2.80E-04$	$1.39E-04$	−2.011	0.0442
Slippery and landslide	$4.06E-01$	$2.05E-01$	1.987	0.0469
Suliflúxion	$3.61E-01$	$1.81E-01$	1.997	0.0458
Temperate dry	$4.93E-01$	$2.22E-01$	2.221	0.0263
Temperate humid	$2.88E-01$	$2.11E-01$	1.363	0.17
Hot and temperate	$1.88E-01$	$2.22E-01$	0.846	0.39
Hot and temperate humid	$8.46E-01$	$3.73E-01$	2.264	0.0235
Flooding area	$1.77E-01$	$1.48E-01$	1.196	0.23
Volcano area	$6.42E-02$	$1.50E-01$	0.428	0.66
SPATIAL CORR. ($\mathbf{y}^*$)	$-1.35E-03$	$3.51E-04$	−3.851	0.0001
Rural or urban	$5.74E-02$	$6.01E-02$	0.955	0.33
People < 5 km from road	$-2.34E-06$	$8.70E-07$	−2.692	0.0071
People 5-15 km from road	$1.95E-06$	$3.00E-06$	0.651	0.51
People > 15 km from road	$-8.83E-07$	$1.20E-05$	−0.074	0.94
County surface (km^2)	$-3.15E-05$	$7.60E-06$	−4.142	$3.45E-05$
Cereal production coeff.	$3.83E-04$	$2.05E-04$	1.871	0.061
Protected area	$1.01E-01$	$7.59E-02$	1.332	0.18
> 35% of irrigation area	$-2.18E-01$	$7.44E-02$	−2.925	0.0034
Closed forest	$3.06E-02$	$6.90E-02$	0.443	0.65
Arable land (30-60%)	$3.01E-02$	$7.61E-02$	0.395	0.69
Arable land (> 60%)	$3.35E-01$	$2.15E-01$	1.556	0.11

the standard errors, too. Even if they do not have a theoretical meaning, they can have a descriptive value and provide some general information. However, some studies where the standard errors are computed both by standard statistical packages and bootstrap simulation techniques remark on the comparability of the results.

In both rural and urban areas, the household variables that have a relationship to a household's probability of being poor are the adult literacy rates (if the components of the household have a diploma, the household's probability of being poor decreases). Environmental factors also show relationship to a household's probability of being poor. In particular, households living close to roads, in large counties and with irrigation systems have a low probability of being poor. It is important to underline the effect of the spatial correlation variable ($\mathbf{y}^*$) that denotes the presence of clusters in the spatial distribution of poverty and the influence among neighbour households on the probability of being poor.

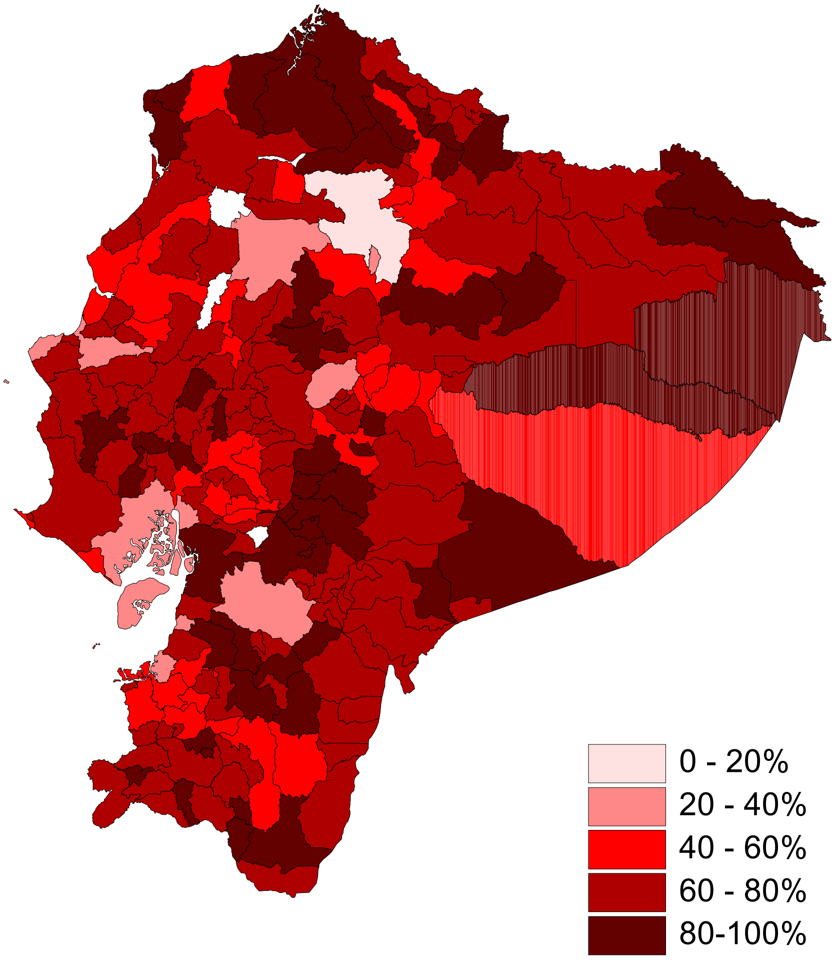

Figure 1: Percentage of poor people in each county.

The estimated parameters of the autologistic model (Equation 2.1) were applied to the data in the county database (INFOPLAN) in order to predict the distribution of poverty across all the counties in Ecuador. The percentage of poor households in each county was obtained from Equation 2.3. In order to count the number of poor individuals, the average households' components in each county were multiplied by the number of poor families in each county (Figure 1). Figure 2 shows the aggregate poverty situation at the region level.

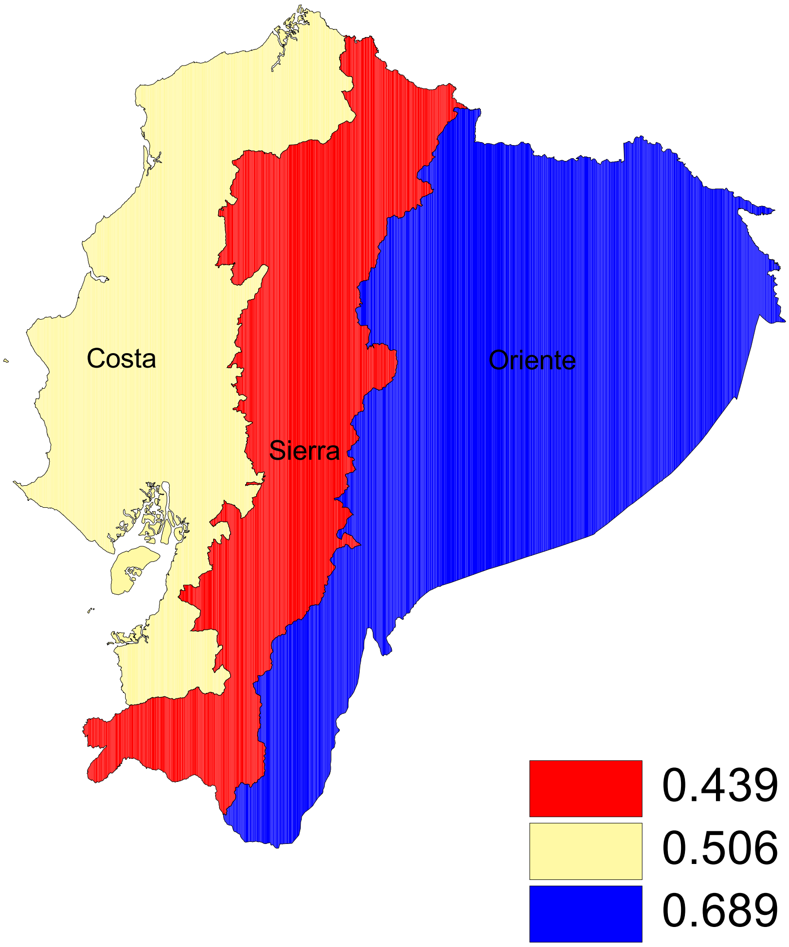

Figure 2: Percentage of poor people in each region.

5 Final remarks

There are some important implications of applying spatial statistical analysis to poverty-mapping studies.

First of all is the importance of taking the spatial dimension of the data into account. After having found a significant spatial correlation between the units, ignoring the spatial component in the regression analysis could lead to misleading estimates of the parameters. This may result in a large proportion of poor households being excluded from the allocation of transfers while a number of non-poor households might be declared potential beneficiaries.

Moreover, the use of spatial models in combination with the visual nature of the poverty maps, obtained from applying the spatial regression methods, may highlight unexpected relationships that would escape notice in a standard regression analysis.

But one of the most important difficulties encountered in spatial analysis concerns the availability of adequate data.

Another problem, which arises in all polygon-based spatial analysis, including the present study, is the modifiable areal unit problem (MAUP): the areal units (administrative or political boundaries, agro-ecological zones, etc.) are arbitrary groupings and the data within each can be aggregated in an infinite number of ways (Nelson, 2001; Bigman and Deichmann, 2000). The implication is that different kinds of aggregation can lead to different results in the spatial analysis so that variables, parameters and processes that are important at one scale or unit are frequently not important or predictive at another scale or unit. A definitive solution to minimize this effect remains to be found.

Poverty can be evaluated using: economic measures such as monetary indicators of households well-being (expenditures, income, consumption, etc.); demographic indicators (gender and age of head of the household, household size, infant mortality rates); and environmental and health measures (access to safe water and sanitation, time spent by household to collect water, cereal production for a family, prevalence of acute infections, disability adjusted life years) (Shyamsundar, 2002). The choice of one indicator rather than another usually depends on the availability of the data and on the practical implications in terms of time, costs and technical requirements for constructing the index. The consequence of using a particular index is that different indicators can lead to different results of the analysis, and so to alternative poverty rankings. One solution, albeit time consuming, could be to apply different kinds of indicators to the same analysis and then to compare and evaluate the implications of each index.

Concluding the results of the fitted spatial model demonstrate the importance of environmental variables which suggests the presence of a poverty-environment relationship and hence the impact of environmental factors on the lives of the poor and on poverty reduction efforts. For this reason, environmental indicators could be an improvement for designing and evaluating poverty reduction strategies and they should be introduced into the statistical analysis.

References

[1] Anselin, L. (1992): *Spatial Econometrics: Method and Models.* Boston: Kluwer Academic Publishers.

[2] Arbia, G. and Espa, G. (1996): *Statistica Economica Territoriale.* Padova: CEDAM.

[3] Bailey, T.C. and Gatrell, A.C. (1995): *Interactive Spatial Data Analysis.* London: Longman.

[4] Besag, J.E. (1974): Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, **35**, 192-236.

[5] Besag, J.E. (1975): Statistical analysis of non-lattice data. *Statistician*, **24**, 179-195.

[6] Besag, J.E. (1977): Efficiency of pseudo-likelihood estimation for simple Gaussian fields. *Biometrika*, **64**, 616-618.

[7] Bigman, D. and Deichmann, U. (2000): Geographic targeting for poverty alleviation. In D. Bigman & H. Fofack (Eds.): *Geographic Targeting for Poverty Alleviation. Methodology and application.* World Bank Regional and Sectoral Studies.

[8] Bigman, D., Dercon, S., Guillaume, D., and Lambotte, M. (2000): Community targeting for poverty reduction in Burkina Faso. *World Bank Econ. Rev.*, **14**, 167-193.

[9] Cliff, A.D. and Ord, J.K. (1981): *Spatial Processes. Models & Applications.* London: Pion Limited.

[10] Cressie, N. (1993): *Statistics for Spatial Data.* New York: John Wiley & Sons.

[11] Haining, R. (1990): *Spatial Data Analysis in the Social and Environmental Sciences.* Cambridge: Cambridge University press.

[12] Hentschel, J., Lanjouw, J.O., Lanjouw, P., and Poggi, J. (2000): Combining census and survey data to trace the spatial dimensions of poverty: a case study of Ecuador. *World Bank Econ. Rev.*, **14**, 147-165.

[13] Nelson, A. (2001): Analyzing data across geographic scales in Honduras: detecting levels of organization within systems. *Ag. Ecosys. Env.*.

[14] O'Neil, R.V., Krummel, J.R. et al (1988): Indices of landscape pattern. *Land. Eco.*, **1**, 153-162.

[15] Openshaw, S. and Taylor, P. (1981): The modifiable unit problem. In N. Wrigley (Ed.): *Quantitative Geography.* London: Pion, 127–144.

[16] Shyamsundar, P. (2002): *Poverty Environment Indicators.* World Bank Environment Department, Paper No. 84.

[17] Strauss, D. and Ikeda, M. (1990): Pseudo-likelihood estimation for social networks. *Journal Am. Stat. Ass.*, **85**, 204-212.