Draft for Comments
May 2008

# An Empirical Evaluation of Poverty Mapping Methodology: A Non-Spatial versus Spatial Approach

Susan Olivia[1]*, John Gibson[2], Scott Rozelle[3], Jikun Huang[4] and Xiangzheng Deng[4]

## Abstract

Poverty maps provide information on the spatial distribution of welfare and can predict poverty levels for small geographic units like counties and townships. Typically regression methods are used to estimate coefficients from the detailed information in household surveys, which are then applied to the more extensive coverage of a census. One problem with standard regression techniques is that they do not take into account the 'spatial dependencies' that often exist in the data. Ignoring spatial autocorrelation in the regression providing the coefficient estimates could lead to misleading predictions of poverty, and estimates of standard errors. Household survey data usually lack exact measures of location so it is not possible to fully account for this spatial autocorrelation. In this paper, we use data from Shaanxi, China with exact measures of distance between each household. A variety of spatial regression models are applied to these data, with the results used as a benchmark for evaluating any bias and inferential errors in poverty mapping regressions that do not incorporate this spatial information.

**JEL: C31, C42, O53, P36**
**Keywords:** China, Poverty, Small Area Estimation, Survey Methods, Spatial Models

[*] Corresponding author: olivia@primal.ucdavis.edu. [1] University of California, Davis; [2] University of Waikato, New Zealand; [3] Stanford University; [4] Chinese Academy of Sciences.

**Acknowledgments:** We are grateful to the Research and Education Advanced Network New Zealand Capability Fund for financial support and to Albert Park, Hongbin Li and Shuming Bao for assistance. All remaining errors are our responsibility.

## 1. Introduction

Most analyses of poverty and inequality in developing countries are based on household surveys that collect detailed information on income or consumption. Because these surveys are costly to implement, the samples are generally limited to a few thousand households. Consequently, poverty and inequality estimates must occur at a high level of aggregation, such as the national or the first sub-national level (e.g. province or region). For example, China's rural household survey samples 80,000 households but yields poverty estimates that are representative only for each province *(n = 31)*. Census data on the other hand, have the required sample size and can provide reliable estimates at highly disaggregated levels but lacks detail on income or consumption which are needed for measuring poverty and inequality.

To fill the gap, poverty analysts have recently experimented with techniques for combining the detailed information from household surveys with the more extensive coverage of census data (*aka* small area estimation). The methodology is developed by Elbers, Lanjouw and Lanjouw (2003), hereafter denoted as ELL, and has been applied to a substantial number of countries, including Albania, Brazil, Cambodia, Indonesia, Madagascar, Mexico, Morocco, and South Africa. In some cases, the poverty maps are used by governments to target financial resources to particularly needy areas[1]. In the approach introduced by ELL, household survey data are used to estimate a model of consumption, with the explanatory variables restricted to those that are also available from a recent census. The coefficients from this estimated model are then combined with the overlapping variables from the census (which cover all households), and consumption and income levels are predicted for each household in the census. Using such data, we can then predict the odds of being poor for each census household and add these up to yield estimated poverty rates for disaggregated (small) geographic units. These welfare indicators are then plotted on a map, which is conventionally called a poverty map.

[1] See http://www.worldbank.org/poverty for a list of applications that apply ELL's (2003) poverty mapping technique.

In a recent report of an expert review panel entitled "Evaluation of World Bank Research, 1998 – 2005", the ELL's (2003) poverty mapping methodology received severe critical comments (Banerjee et. al., 2006). The report claims that while the method has been a popular tool in many countries in recent years, it is increasingly understood that there are problems with the methods, or at the very least better ways to improve the precision of the predictions than the methods typically used. Above all, the major shortcoming of using ELL technique to generate poverty maps is that it does not take into account strong *spatial autocorrelation* that have been found in many data sets. This *spatial autocorrelation* can arise either because nearby locations have unobserved factors in common (e.g. deteriorating environmental conditions) or because of interaction between one household and another (e.g. poverty rate in one area is directly affected by poverty in nearby areas). The first model, of unobserved common factors, is known as a *spatial error* model while the second, of neighbour's interactions, is a *spatial lag* model. If this autocorrelation is ignored, the calculated standard errors will overstate the true precision of the estimates (Tarozzi and Deaton, 2007). This misleading sense of precision may cause policy makers to target particular areas which in reality are no poorer than other areas that do not get targeted. In response to this criticism, Elbers et al. (2008) point out that one may reduce the impact of the correlations to negligible levels by introducing a variety of cluster means calculated from the census or from a tertiary data set such as GIS data into the first stage model. They also note that the ELL methodology can deal with autocorrelation problem by redefining the clusters at the broader level and rerun the analysis using this redefined cluster ("conservative approach", p.29).

While the ELL poverty mapping technique attempts to deal with spatial autocorrelation, it necessarily does so in a way which does not rely on knowing the location of either sample or census households. In this paper we use ex-post geo-referenced household income and expenditure survey data from rural Shaanxi, China that allow exact distances between each household to be measured as well as information from the 2000 Population Census and a rich set of GIS-linked environmental variables derived from high resolutions satellite imagery to estimate poverty and inequality for small areas in rural Shaanxi. This additional information on distance between neighbours inside a cluster, and distance to neighbours in other clusters, allows for more explicit modelling of spatial autocorrelation. The aim of such modelling is to see whether inferences based on the current poverty mapping methodology are the same as those which would be reached if researchers knew where households were located and were able to more properly model spatial effects. We apply two different methods to get welfare estimates for small areas in Shaanxi province, China: the non-spatial (ELL) method and new methods that use spatial econometrics approaches (which has not been adequately considered in previous studies) to account for unobserved spatial correlations. We compare the results from each method and assess how well the use of new data and/or new estimation procedures would improve the effectiveness of analyses that explore such spatial dimensions of poverty. These comparisons may matter since there are unpleasant consequences of modelling spatial effects in the wrong way. For example, ignoring a spatial error structure can cause inference problems while ignoring spatial lags can bias coefficient estimates since the omitted autocorrelation in the lag model enters through the systematic part of the model (Anselin, 1988). In this paper we also investigate whether using a rich set of GIS linked geophysical variables is more effective in dealing with spatial autocorrelation than using the census means as advised by Elbers et al. (2008).

The rest of the paper is organised as follows. Section II briefly discusses the non-spatial approach in poverty mapping methodology followed by spatial approach in Section III. In Section IV, we describe the data used in the paper. Section V presents both the non-spatial and spatial results in the first stage model of consumption in the poverty mapping methodology. Section V concludes.

## 2. The ELL Methodology

The non-spatial approach will follow Elbers et al. (2003), in which the econometric analysis consists of two stages. In the first stage, a model of (log) per capita consumption expenditure $y_i$ is estimated:

$$\ln\ y_i = \mathbf{x}_i\boldsymbol{\beta} + u_i \qquad (1)$$

where $\mathbf{x}_i$ is the vector of explanatory variables for the $i$th household and is restricted to those variables that can also be found in the census, $\boldsymbol{\beta}$ is a vector of parameters and $u_i$ is the error term satisfying $E[u_i \mid x_i] = 0$. This error term can be decomposed into two independent components: a cluster specific effect $\eta_c$ and a household specific effect $\varepsilon_{ci}$. This complex error structure allows for both spatial autocorrelation (that is, a 'location effect' common to all households in the same area) and heteroskedasticity (non-constant variance) in the household component of the error term.

In the second stage of the analysis, the estimated regression coefficients from equation (1) are applied to data from the 2000 Population Census using the characteristics included in the vector $\mathbf{x}_i$ to obtain predicted consumption for each household within the micro census. While it is possible to directly predict consumption by simply combining the characteristics for census household $j$, $\mathbf{x}_j^c$ with $\hat{\boldsymbol{\beta}}$ from equation (1), a more refined methodology is needed to account for the complex nature of the disturbance term (Elbers et al., 2003). Specifically, estimates of the distribution for both $\eta$ and $\varepsilon$ are obtained from the residuals of equation (1) and from an auxiliary equation that explains the heteroskedasticity in the household-specific part of the residual. Following Elbers et al. (2003), the auxiliary equation is estimated using a logistic model of the variance of $\varepsilon_{ci}$ conditional on $\mathbf{z}_{ci}$:

$$\ln\left[\frac{\varepsilon_{ci}^2}{A-\varepsilon_{ci}^2}\right] = \mathbf{z}_{ci}'\hat{\alpha} + r_{ci} \qquad (2)$$

where $\mathbf{z}_{ci}$ is a set of potential variables that best explain the variations in $\varepsilon_{ci}^2$, and $A$ is set equal to $1.05\times\max\{\varepsilon_{ci}^2\}$. In this stage, we also conduct a series of simulations, and for each simulation, we draw a set of beta and alpha coefficients, $\tilde{\boldsymbol{\beta}}$ and $\tilde{\alpha}$, from the multivariate normal distributions described by the first stage point estimates and their associated variance-covariance matrices. Additionally, we draw $\tilde{\sigma}_\eta^2$, a simulated value of the variance of the location error component. Combining the alpha coefficients with census data, for each census household we estimate $\tilde{\sigma}_{\varepsilon,ci}^2$, the household-specific variance of the household error component. Then for each household we draw simulated disturbance terms, $\tilde{\eta}_c$ and $\tilde{\varepsilon}_{ci}$ from their corresponding distributions. We simulate a value of expenditure for each household, $\hat{y}_j^c$ based on both predicted log expenditure, $\mathbf{x}_j'^c \tilde{\boldsymbol{\beta}}$ and the disturbance terms:

$$\hat{y}_j^c = \exp\left(\mathbf{x}_j'^c \tilde{\boldsymbol{\beta}} + \tilde{\eta}_c + \tilde{\varepsilon}_{ci}\right) \tag{3}$$

Finally, the full set of simulated $\hat{y}_j^c$ values are used to calculate expected values of distributional statistics, including poverty measures for each 'local area' and for higher level aggregations of local areas. We repeat this procedure 100 times, drawing a new set of coefficients and disturbance terms for each simulation. For any given location (such as a county or township), the mean across the 100 simulations for a given statistic such as the headcount poverty rate, provides the point estimate of those statistics for that location, while the standard deviation serves as an estimate of the standard error.

**3. A spatial regression approach**

As discussed earlier, a major weaknesses of conventional statistical methods used to produce poverty maps is that they do not take into account strong dependencies that may be correlated with poverty and tend to occur in clusters of villages at the same time. Thus, if there is a significant spatial correlation among the households within a village due to some real but some unobserved factor, then ignoring the spatial component in the regression analysis could lead to misleading estimates of the parameters (Anselin, 1988). If this were the case, such analysis could result in a large proportion of poor households being excluded say from the allocation of transfers, while a number of non-poor households might be deemed as potential beneficiaries. Thus, to analyze the distribution of poverty more accurately, novel methods which use new spatial data and analytical tools would help to improve the effectiveness of analyses that explore such spatial dimensions of poverty.

A *spatial weight matrix W,* is one way of imposing the required structure on the study of spatial autocorrelation. This is an $N \times N$ positive and symmetric matrix which exogenously determines for each observation (row) which locations (columns) belong in its neighborhood. For non-neighbors, $w_{ij}$=0, while for neighbors the weights are either $w_{ij}$=1 (binary weights) or a function of something else, such as: $w_{ij} = 1/d_{ij}$ where $d_{ij}$ is the distance between observations $i$ and $j$ (inverse distance weights). Who is a neighbor may be defined either by a distance criteria, especially with point data, or by whether they share a common border and/or vertex (contiguity) for areal data (Wilhelmsson, 2002). The diagonal elements of the weights matrix are conventionally set to zero, and typically standardized such that the elements of a row sum to one (Anselin and Bera, 1998). Hence, the spatial weight matrix allows all of the interactions between observation $i$ and each of its neighbors to be parameterized in the form of a weighted average. Specifically, for some random variable of interest $z$, each element of the spatially lagged variable $Wz$ equals $\sum_j w_{ij} z_j$ which is a weighted average of the $z$ values in the neighborhood of point $i$.

According to Anselin (1988), there are two major ways in which spatial autocorrelation can manifest itself: spatial lag dependence and spatial error dependence. This provides the theoretical basis for a so called spatial lag model (spatial autoregressive model) and spatial error model. Spatial lag dependence refers to a situation in which the dependent variable in one area is affected by the dependent variable in nearby areas. For instance, if the dependent variable is income or poverty, it is likely that the level of economic activity in one area is directly affected by the level of economic activity in neighboring areas through migration or trade-investment linkages. If the regression analysis is carried out without adjustment for spatial lag dependence, the estimate coefficients will be biased and inconsistent.

Formally, the spatial lag model is defined as:

$$Y = \rho WY + X\beta + \varepsilon \tag{4}$$

where $Y$ is an $N \times 1$ vector of observations on the dependent variable, $WY$ is the spatially lagged dependent variable, $X$ is an $N \times k$ matrix of explanatory variables, $\varepsilon$ is a vector of errors, $\beta$ is the vector of regression parameters and $\rho$ is the spatial autoregressive parameter. Although equation (4) looks like a dynamic model from time-series analysis, one key difference causes OLS (the conventional regression model) to always be an inconsistent estimator of the spatial lag model. In the time-series context, if there is no serial correlation in the errors, $\varepsilon_t$ there will be no correlation between $y_{t-1}$ and $\varepsilon_t$ and OLS will be a consistent estimator. In contrast, $(WY)_i$ is always correlated with both $\varepsilon_i$ and the error term at all other locations. Hence, OLS is not consistent for the spatial lag model (Anselin, 1988).

The second manifestation is through spatial error dependence. In this situation, the error term in one area is correlated with the error term in nearby areas. This can happen if there are variables that are not included in the regression model but do have an effect on the dependent variable (omitted variable bias problem) and they are spatially correlated. For example, the quality of local government and environment factors affects income and poverty, but it is difficult to include in a regression model. Because the quality of local government and environment is likely to be spatially correlated, the error term in each area is likely to be correlated with those in nearby areas. This consequently violates one of the underlying assumptions of the OLS regression model that the disturbance terms for each observation are not correlated with one another. In this case, the estimates of the coefficient are no longer efficient and the standard $t$ and $F$ tests will produce misleading inference (Anselin, 1988).

In contrast to the spatial lag model, the spatial error model is defined as:

$$\begin{aligned} Y &= X\beta + \varepsilon \\ \varepsilon &= \lambda W\varepsilon + \mu \end{aligned} \qquad (5)$$

where $\lambda$ is the spatial autoregressive coefficient, $\mu$ is a vector of errors that are assumed to be independently and identically distributed and the other variables and parameters are as defined in equation (4). In this model, the error for one observation depends on a weighted average of the errors for neighboring observations, with $\lambda$ measuring the strength of this relationship.

It is clear that both equations (4) and (6) are restricted versions of a more general spatial autoregressive model (SAC) with autoregressive disturbances:

$$\begin{aligned} Y &= \rho W_1 Y + X\beta + \varepsilon \\ \varepsilon &= \lambda W_2 \varepsilon + u \end{aligned} \tag{7}$$

It may therefore seem preferable to always begin with a model like equation (7) and test in a general-to-specific way to see if either equation (6) or equation (4) are data-acceptable. Indeed, equation (7) could always be the starting point for cross-sectional regressions because the standard OLS regression model:

$$Y = X\beta + \varepsilon \tag{8}$$

is just a special case with $\rho=\lambda=0$. However, spatial models are much more computationally demanding and for most statistical software there are limits on the sample sizes that they can accommodate. Moreover, they have to be estimated by methods such as instrumental variables and maximum likelihood that require additional assumptions (Anselin, 1988).

In this paper, we experiment with global spatial regression analysis (i.e. the model assume that the relationship between poverty and geographic variables is the same across the country) to analysis of poverty. As discussed earlier, when spatial autocorrelation is present in the data, the OLS result may be biased, and standard errors are both biased and inefficient, leading to invalid inferences. To control for it, we will first detect the spatial autocorrelation using the standard global statistics that have been developed such as Lagrange Multiplier (LM), which only need the restricted model to be estimated. Therefore it is common in the spatial econometrics literature to start with an OLS model and use the residuals from that model to test against spatial alternatives. In addition to these LM tests, Moran's $I$ test, which has some parallels with the Durbin-Watson statistic, is also widely used (Anselin and Bera, 1998). For a row-standardized spatial weight matrix, Moran's $I$ can be expressed as:

$$I = \frac{\mathbf{e}'\mathbf{W}\mathbf{e}}{\mathbf{e}'\mathbf{e}} \tag{9}$$

where **e** is a vector of OLS residuals and **W** is the spatial weight matrix. Moran's $I$ is asymptotically normally distributed with mean $-1/(N-1)$ and its statistical significance can be evaluated from a standardized normal table. A feature of Moran's $I$ is that the alternative hypothesis does not specify the process generating the autocorrelated disturbances. However, there is a simple intuition for Moran's $I$ because for any variable **z** in deviation from mean form, $I$ is equivalent to the slope coefficient in a linear regression of **Wz** on **z** (Anselin, 1995).

The LM tests are based on explicitly specified alternative hypotheses. For testing OLS against the spatial error model ($\lambda$=0) the test statistic is:

$$LM_{\lambda} = \left[\mathbf{e}'\mathbf{W}\mathbf{e}/\hat{\sigma}^{2}\right]^{2}/T \tag{10}$$

where $T = tr(W' + W)W$ and $LM_{\lambda}$ is distributed as $\chi^2$ with 1 degree of freedom. For testing OLS against the spatial lag model ($\rho$=0) the test statistic is:

$$LM_{\rho} = \left[\mathbf{e}'\mathbf{W}\mathbf{Y}/\hat{\sigma}^{2}\right]^{2}/T_{1} \tag{11}$$

where $T_1 = (\mathbf{WX}\hat{\beta})'\mathbf{M}(\mathbf{WX}\hat{\beta})/\hat{\sigma}^2 + T$ and $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. One difficulty with both $LM_{\lambda}$ and $LM_{\rho}$ is that they each have power against the other alternative. In other words, when testing $\lambda$=0, $LM_{\lambda}$ responds to nonzero $\rho$ and when testing $\rho$=0, $LM_{\rho}$ responds to nonzero $\lambda$. To test in the possible presence of both spatial error and spatial lags, Anselin et al. (1996) develop specification tests for spatial lags that are robust to ignored spatial errors and tests for spatial errors that are robust to ignored spatial lags. These tests denoted $LM^{*}_{\lambda}$ and $LM^{*}_{\rho}$ should be used when both $LM_{\lambda}$ and $LM_{\rho}$ are statistically significant.

All five of the spatial autocorrelation tests described here will be carried out. Depending on the outcome of the specification tests, the regression model for the first stage consumption equation will be re-estimated in either the spatial lag or spatial error framework.

**4. Data**

The data come from three sources: (i) the 2000 Population Census; (ii) the 2001 Rural Household and Income Expenditure Survey conducted by the China's National Bureau of Statistics; and (iii) satellite remote sensing for Shaanxi. Shaanxi is selected because it is an area of high poverty in China, with an incidence of poverty in 2000 that was 2.9 times as high as national average. Furthermore, it has one of the slowest rates of poverty reduction in China since 1981 (Ravallion and Chen, 2007). Table 1 indicates which variables come from each of these three sources, distinguishing between those available for the sample and those available for the population. The methodology, which will be discussed below, requires the model of consumption to be estimated on the sample observations and the coefficients then applied to population data on the same variables. Table 1 also presents the mean values of the explanatory variables available in both the household survey and the population census that were selected for inclusion in the model of consumption.

The latest population census was conducted in November 2000. Like the census in many other countries, the Chinese version did not collect information on income and expenditure, however it provides information on a number of characteristics that are likely to be correlated with consumption and poverty. It includes information on demographics, education, economic activities and the attributes of the dwelling. We use a 1 percent sample of the census (henceforth, a micro-census), which was designed to be representative at the township level. The census listed 2,144 townships and almost 76,000 rural households from these townships are listed in the micro-census.

The 2001 Rural Household Income and Expenditure Survey (RHIES), as its name implies, collected information on the income and expenditure of households. Apart from this, the survey also collected information on household characteristics, employment, seasonal labor migration, agricultural production, dwelling characteristics, ownership of durable goods and fixed assets and access to public infrastructure. The RHIES used a random multi stage systematic sampling of 1,400 households in Shaanxi. In the first stage, 25 counties were selected, in which between 4 – 8 townships were selected from each county. From each township, 1 village was selected and 10 households were selected from each selected village.

Despite the RHIES collecting high quality data on people's living standards, it is sample and is small relative to population that it is trying to represent. Figure 1 shows that 25 (128) sampled counties (townships) were selected from among the 107 (2144) counties (townships) in the province[2]. The sample size in this survey is therefore too small to allow an estimation of the incidence of poverty at say the county or township level. As a result, poverty estimates from this source of data must occur at a high level of aggregation, such as province or possibly prefecture level.

Like the household survey in many other developing countries, the Chinese version did not geo-reference households in part because of lack of information about the benefits. In a recent paper, Gibson and McKenzie (2007) argue that the data collection of GPS coordinates should become a routine part of household surveys and census, since doing so can lead to better economics and policy advice. To assess whether knowing the position of households relative to each other can improve the modeling of spatial autocorrelation, one of the authors lead an expedition team to Shaanxi in November 2007 to record the locations of each households in the 2001 RHIES using a Global Positioning System (GPS) receiver. Figure 2 shows the location of the households being surveyed in the 2001 RHIES.

The environmental component of this research uses a variety of spatially referenced variables that provides information on land cover, rainfall, temperature, elevation and terrain slope for Shaanxi, which can be considered part of what Ravallion (1998) calls geographic capital. The land cover data are from satellite remote sensing data provided by the US Landsat TM/ETM images which have a spatial resolution of 30 by 30 meters. These data have been interpreted, involving considerable ground-truthing and aggregated

[2] In the context of China, administrative levels start from the national level, go down to province (*sheng*), prefecture (*di qu*), county (*xian*) and township (*xiang).*

into 1 kilometer by 1 kilometer at the county level by Chinese Academy of Sciences – CAS (Liu et al., 2003a and 2003b). These data have previously been used by Deng et al. (2002, 2003 and 2008). A hierarchical classification system of 25 land-cover classes was applied to the data and the total land area of each county were aggregated from the 25 classes of land cover in this study. The data for measuring *rainfall* (measured in millimetres per year) and *temperature* (measured in degrees centigrade per year) are from the CAS data centre but were initially collected and organized by the Meteorological Observation Bureau of China from more than 600 national climatic and meteorological data centres.

The *elevation* and *terrain slope* variables, which measure the nature of the terrain of each county, are generated from China's digital elevation model data set that are part of the basic CAS data base. A variable to measure the share of plain area is also is created by dividing the land area in a county that has a slope that is less than eight degrees by the total land area of the county. Information on the properties of soil also is part of our set of geographic and climatic variables from the CAS data center. Originally collected by a special nationwide research and documentation project (the *Second Round of China's National Soil Survey*) organized by the State Council and run by a consortium of universities, research institutes and soils extension centres, we use the data to specify two variables: the loam and organic content of the soil (measured in percent).

In addition, a variable that measures the density of a county's highway network is also included in this study. This variable is based on a digital map of transportation networks that exist in each county. It was developed by CAS and the measure includes all highways, national expressways, provincial-level roads and other more minor roads in the mid-1990s. The variable (henceforth—highway density) is measured as the total length of all highways in a county divided by the land size of the county.

## 5. Results

### 5.1. Non-spatial Results

The results of estimating the first stage model of consumption with OLS for 1,070 rural sampled households in Shaanxi are reported in Table 2. This estimator does not take account for any spatial autocorrelation, in common with the recent poverty mapping applications. The first model reported in column 1 of Table 2 is for the regression just on the household characteristics. The model suggests that per capita consumption is higher for households with larger dwellings (as a proxy for housing quality and wealth), with a greater number of their members engaged in the non-agricultural sector. Having access to safe drinking water as well as having sanitary facility in the house also leads to a higher level of consumption. On the other hand, consumption is lower for households with a greater proportion of kids aged 6 years and below, greater proportion of youths aged 7 – 15 years, greater proportion of adults and greater proportion of elderly in the household. An important point to note about these results is that none of these relationships should be treated as causal since the purpose of the first stage model is just to have the best prediction model of consumption.

To capture excluded location effects and other elements of geographical capital, we augment the model in column (1) with environmental variables. Inclusion of environmental variables raises the value of $R^2$ of the consumption model from 0.21 to 0.27 and these variables are jointly statistically significant with a *F-statistic* of 9.65, suggesting that consumption is highly related to the characteristics of the environment of where people live. The environmental variables show that consumption is lower for households in areas on steep slopes, with higher temperature and soils with higher percentage of organic matter. Soils with lower percentage of loam, lower annual rainfall, are all correlated with lower consumption. On the other hand, consumption is higher for households in areas with higher total area of land and higher density of highways.

Column (3) of Table 2 reports the first stage model of consumption based on household characteristics as well as the township level means of the household level variables from the census. The use of census means in the survey model of consumption has been recommended by Elbers et al. (2003) as a way to proxy for location-specific correlates of consumption, which can help to make the cluster specific variance $\eta_c$ smaller and improve precision of the second stage predictions. This model has an $R^2$ of 0.25, as compared with 0.21 for the model in column (1) that is without the census means but otherwise has the same variables. However, many of the added variables are statistically insignificant. The coefficients on most of the household variables that were already in the model generally maintain their size and significance.

The last column of Table 2 reports the results of augmenting the model with household characteristics with environmental variables and means of the census variables. Most of the household characteristics in this model maintain the same sign as they had in the model estimated only on household variables (i.e. the model reported in column (1)). However, the inclusion of the location variables (both environmental and census means) reduces the size and significance of the coefficient on housing area and dummy for households with access to safe drinking water, increases the significance of the coefficients on household members with college degree and engaged in non-agricultural activities as well as households having sanitary facility in the house. The inclusion of the township means of the census variables also alters the significance of the coefficients on log density of highway, percentage loam in the soil, log annual rainfall, percentage of organic matter, temperature and percentage of plain area.

Although most of the explanatory variables reported in Table 2 are statistically significant and have signs that are accord with expectations, however, such conclusions may be premature because the OLS estimates do not account for spatial autocorrelation. To test for spatial autocorrelation, a spatial weighting matrix is needed and in turn this requires a measure of distance between households. Latitude and longitude coordinates for each household were used to calculate this and the weighting matrix is based on inverse distance weights and a neighbourhood size of seventeen kilometres (i.e. the minimum feasible neighbourhood to prevent “islands” with no neighbours).

To test for spatial autocorrelation in the OLS residuals of the consumption equation, both Moran's *I* statistic and several Lagrange Multiplier (LM) tests were used. These LM tests help choose between the two models that can cause spatial autocorrelated residuals: the spatial lag model and the spatial error model. With a spatial lag model, the consumption of each household is affected by the spatially weighted average of consumption of nearby households – even after controlling for observable factors that might be common for the households. In contrast to the spatial lag model, the spatial error model is based on the assumption that spatially-varying omitted factors show up in the model's disturbances, causing the disturbance for one observation to be correlated with a spatially weighted average of neighbouring disturbances.

Spatial autocorrelation diagnostics show that this type of spatial dependence is indeed present (Table 3). The Moran I statistic is statistically significant at 1 percent level. This is further confirmed by the significance of Lagrange Multiplier tests of spatial autocorrelation. This causes problem for the OLS estimator as the presence of correlated errors violates the Gauss-Markov assumption of uncorrelated random errors and more broadly the assumption of independence between observations. This indicates that the OLS is misspecified and that spatial effects should be included in the model. The much larger Lagrange Multiplier in the spatial error model indicates that this type of spatial dependence is more likely in the first stage of consumption model for rural households in Shaanxi.

**5.2. Spatial Regression Results**

Table 3 shows the results of regressing per capita consumption on three types of unrestricted exogenous variables (household characteristics, environmental variables and census means of household characteristics) using the general spatial model (SAC) as shown in equation (7). The first stage model of consumption explains from 23.3 percent of the variation in the model when basing just on household characteristics to 29.4 percent when the model is augmented with both the environmental variables and census means of household characteristics. As can be seen from the table, the spatial parameter corresponding to the spatial lag model ( $r$ ) is statistically significant across 4 sets of estimation. The value of the correlation coefficient using only the household characteristics $\rho = 0.59$ indicates that on average, a 10 percentage point increase in per capita consumption in a particular location will result in a 5.9 percentage point increase in the per capita consumption in a neighbouring location *ceteris paribus*. This seems to suggest a strong evidence of spill over effects in rural Shaanxi.

The results reported in Table 4 also allows one to assess which set of augmenting variables: census means or environmental variables have the most impact in soaking up unwanted spatial autocorrelation in poverty mapping methodology. According to our results, including environmental variables in the first stage model of consumption greatly reduces the spatial correlation by almost 60 percent ( $\rho_{hh} = 0.59$ v.s. $\rho_{hh+env} = 0.24$ ). On the other hand, the reduction is rather minute (about 5.7 percent) if one augments the model with census means of household characteristics. This result suggesting that including environmental variables can be more effective to reduce the impact of the correlations to negligible levels in comparison to the advice suggested by ELL (2003) by introducing a variety of cluster means calculated from the census. Further evidence of this is reported in Table 3, where we fail to reject the null hypothesis of no spatial correlation in the model where both household characteristics and environmental variables are included in the analysis. In other words, the residual of the model with environmental variables does not exhibit spatial autocorrelation compared to the OLS residual based on either household characteristics only or household characteristics augmented with census means.

Tables 5 and 6 contain results of the spatial error and spatial lag models. According to the maximum likelihood estimates, in the preferred spatial error specification $\lambda$*s* are all statistically significant across four different models. The value of $\lambda = 0.125$ with a standard error of 0.01 for the model of per capita consumption based on household characteristics augmented with both environmental variables and census means of household characteristics indicates that the spatially weighted residual of consumption within a 17 kilometre radius is significantly associated with the residual of consumption for a particular household even after controlling for household characteristics and set of location attributes. According to these results, consumption is higher for households with a greater proportion of kids aged 6 years and below, youths aged 7 – 14, adults aged 16-60 and elderly aged 60 years above. On the other hand, households with larger dwelling, having access to safe drinking water, having sanitary facility as well as larger members of households engaged in non-agricultural activities appear richer. When the spatial error model is used, standard errors are generally smaller than for the one in the OLS. The number of significant variables decreased after taking into account for spatial effects. Several of the environmental variables such as elevation and percentage of loam in the soil become statistically insignificant. This could signal the misleading effect spatial error autocorrelation may have on inference using OLS estimates.

**6. Conclusions**

In this paper, we take an explicit spatial econometric approach, which includes testing for the presence of spatial autocorrelation and estimating specifications that incorporate spatial dependence in the first stage of consumption model of the poverty mapping exercises. The alternative spatial econometric models are superior to the OLS estimates by virtue of lower standard errors and free from residual spatial autocorrelation. The significance of the spatial parameters indicates that the OLS model is mis-specified. Our results also seem to suggest that more robust inferences are likely to come from knowing actual distance between households, which in this sense are supportive of the growing use of GPS in household surveys (Gibson and McKenzie, 2007). In addition, we also found that to reduce the effect of spatial correlation in the data by inserting census means alone in the first stage regression model is not as efficient in soaking up unwanted spatial autocorrelation in comparison to include environmental variables into the analysis.

**References**

Anselin, L. (1988). *Spatial Econometrics: Methods and Models.* Kluwer Academic Publishers: Dordrecht.

Anselin, L. (2002). "Under the hood: Issues in the specification and interpretation of spatial regression models", *Agricultural Economics* 27(3): 247 – 267.

Anselin, L., Bera, A., Florax, R. and Yoon, M. (1996). "Simple diagnostic tests for spatial dependence", *Regional Science and Urban Economics*, 26(1): 77 – 104.

Anselin, L. and Bera, A. (1998). "Spatial dependence in linear regression models with an introduction to spatial econometrics" in Ullah, A. and Giles, D. (eds), *Handbook of Applied Economic Statistics.* Marcel Dekker: New York.

Banerjee, A., Deaton, A., Lustig, N. and Rogoff, K. (2006). "An evaluation of World Bank Research, 1998 – 2005" *mimeo* Washington, D.C.: World Bank

Elbers, C., Lanjouw, J., and Lanjouw, P. (2003). "Micro-level estimation of poverty and inequality", *Econometrica* 71(1): 355 – 364.

Elbers, C., Lanjouw, P., and Leitye, P. (2008). "Brazil within Brazil: Testing the poverty map methodology in Minas Gerais", *World Bank Policy Research Working Paper No. 4513.* World Bank: Washington, D.C.

Gibson, J. and McKenzie, D. (2007). "Using global positioning systems in household surveys for better economics and better policy", *World Bank Research Observer* 22(2): 217 – 241.

Hentschel, J., Lanjouw, J., 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 Economic Review* 14(1): 147 – 65.

Tarozzi, A. and Deaton, A. (2007). "Using census and survey data to estimate poverty and inequality for small areas", *mimeo* Department of Economics, Duke University.

**Figure 1.** Sampled Counties and Townships in the Rural Household Income and Expenditure Survey for Shaanxi

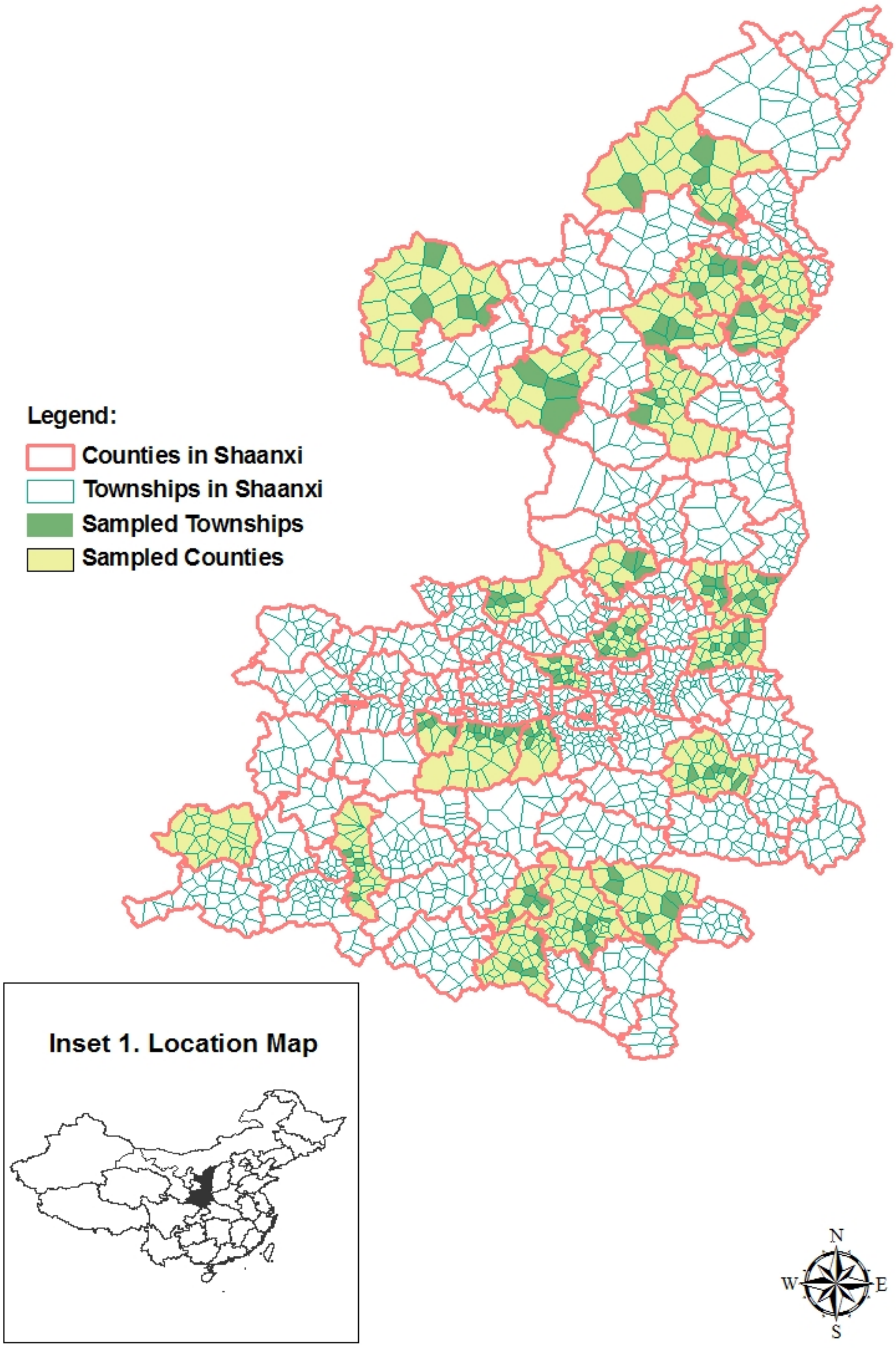

**Figure 2.** Location of the sampled households in Shaanxi

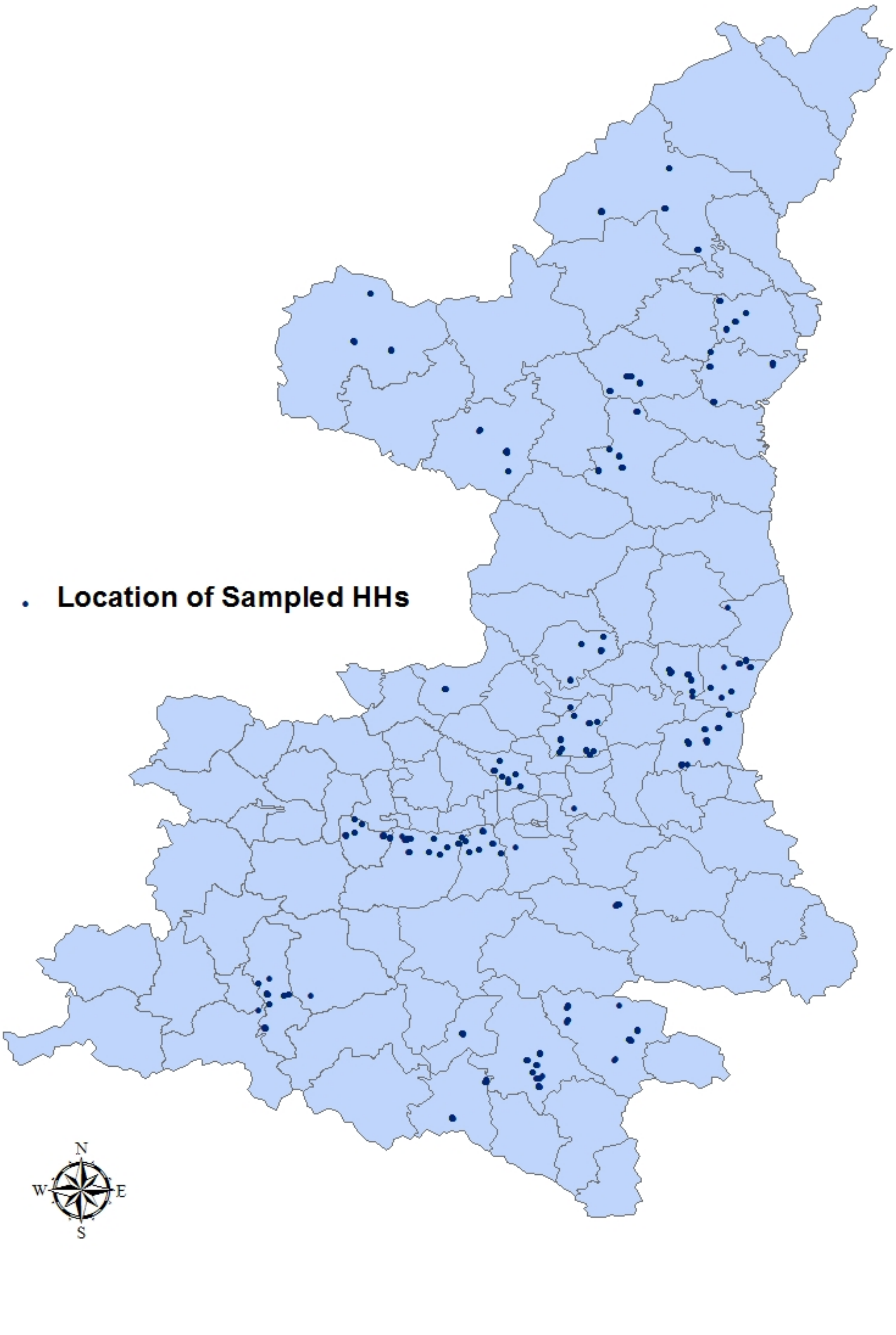

**Table 1**. Availability of data and sources

| | Sample | Survey Mean | Population | Census Mean |
|---|---|---|---|---|
| **Welfare Indicator(s)** | | | | |
| Per capita expenditure | HIES | 1,090.68 | n.a. | n.a. |
| **Demographic Characteristics** | | | | |
| Number of persons aged 6 and below | HIES | 0.24 | Census | 0.29 |
| Number of persons between 7 & 15 years of age | HIES | 0.98 | Census | 0.79 |
| Number of persons between 16 & 60 years of age | HIES | 2.88 | Census | 2.27 |
| Number of persons aged 61 and above | HIES | 0.27 | Census | 0.37 |
| **Education Characteristics** | | | | |
| # of labor force in HH completed primary school | HIES | 0.75 | Census | 0.84 |
| # of labor force in HH completed junior high school | HIES | 1.25 | Census | 1.06 |
| # of labor force in HH completed senior high school | HIES | 0.29 | Census | 0.21 |
| # of labor force in HH completed vocational school | HIES | 0.03 | Census | 0.04 |
| # of labor force in HH with college degree and above | HIES | 0.01 | Census | 0.01 |
| **Dweling Characteristics** | | | | |
| Housing area (in square meter) | HIES | 101.23 | Census | 118.01 |
| Brick house (dummy = 1; 0 otherwise) | HIES | 0.52 | Census | 0.55 |
| Household uses LPG as main source of cooking (dummy = 1; 0 otherwise) | HIES | 0.01 | Census | 0.02 |
| **Household economic activities** | | | | |
| Number of household members engage in non-agriculture activities | HIES | 0.57 | Census | 0.38 |
| **Geophysical variable(s) at county level** | | | | |
| Total areas of land | Geo | 249,641 | Geo | 219,993 |
| Percentage of plain area | Geo | 0.16 | Geo | 0.17 |
| Percentage of loam in the soil | Geo | 0.29 | Geo | 0.30 |
| Percentage of organic matter | Geo | 0.63 | Geo | 0.75 |
| Annual rainfall | Geo | 650.06 | Geo | 681.85 |
| Temperature | Geo | 10.08 | Geo | 10.18 |
| Density of highway in m/1000 ha (log) | Geo | 9.2 | Geo | 11.00 |
| Slope (log) | Geo | 0.99 | Geo | 1.07 |
| Elevation (log) | Geo | 6.83 | Geo | 6.81 |

**Table 2.** First Stage Regression Model of Per Capita Expenditure (OLS)

| Variable | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| ***Household Level Characteristics*** | | | | |
| # HH members age <6 | -0.290*** | -0.341*** | -0.305*** | -0.303* |
| | (0.043) | (0.040) | (0.043) | (0.042) |
| # HH members age 7 - 15 years | -0.095*** | -0.118*** | -0.108*** | -0.103* |
| | (0.023) | (0.023) | (0.023) | (0.023) |
| # HH members age 16 - 60 years | -0.064*** | -0.122*** | -0.078*** | -0.080* |
| | (0.027) | (0.026) | (0.028) | (0.028) |
| # HH members age > 60 years | -0.224** | -0.223*** | -0.219*** | -0.232* |
| | (0.035) | (0.035) | (0.035) | (0.034) |
| # HH members completed primary school | -0.163*** | -0.093*** | -0.142*** | -0.134* |
| | (0.034) | (0.033) | (0.034) | (0.033) |
| # HH members completed junior high school | -0.064** | -0.028 | -0.056* | -0.055* |
| | (0.031) | (0.031) | (0.033) | (0.032) |
| # HH members completed senior high school | 0.005 | 0.063 | -0.004 | 0.006 |
| | (0.044) | (0.044) | (0.045) | (0.044) |
| # HH members completed vocational degree | 0.166 | 0.157 | 0.205 | 0.140 |
| | (0.119) | (0.115) | (0.117) | (0.115) |
| # HH members with college degree and above | 0.238 | 0.291* | 0.224* | 0.261* |
| | (0.178) | (0.173) | (0.177) | (0.173) |
| Housing area (meter square) | 0.003*** | 0.003*** | 0.002 | 0.003 |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| HH uses LPG as main cooking fuel (dummy = 1; 0 otherwise) | 0.398 | 0.230 | 0.391*** | 0.398 |
| | (0.374) | (0.365) | (0.371) | (0.363) |
| House made of brick (dummy = 1; 0 otherwise) | 0.051 | 0.103** | -0.015 | 0.000 |
| | (0.041) | (0.044) | (0.046) | (0.056) |
| HH has access to safe drinking water (dummy = 1; 0 otherwise) | 0.176*** | 0.066 | 0.149*** | 0.077* |
| | (0.041) | (0.044) | (0.043) | (0.044) |
| HH has toilet facility in the house (dummy = 1; 0 otherwise) | 0.191*** | 0.220*** | 0.297*** | 0.284* |
| | (0.048) | (0.054) | (0.052) | (0.062) |
| # HH members engaged in non-agricultural activities | 0.119*** | 0.113*** | 0.131*** | 0.124* |
| | (0.030) | (0.029) | (0.031) | (0.031) |
| ***Environmental Variables*** | | | | |
| Total area of land | | 0.054 | | 0.152* |
| | | (0.054) | | (0.078) |
| Elevation (log) | | 0.172* | | 0.185 |
| | | (0.104) | | (0.112) |
| Density of highway (log) | | 0.035* | | 0.047 |
| | | (0.007) | | (0.008) |
| % loam in the soil | | 0.005*** | | 0.012* |
| | | (0.005) | | (0.006) |

| | | | | |
|---|---|---|---|---|
| Annual rainfall (log) | | 0.660*** | | 0.809* |
| | | (0.120) | | (0.170) |
| Slope (log) | | -0.163*** | | -0.103* |
| | | (0.054) | | (0.063) |
| % organic matter in soil texture | | -0.422** | | -0.612 |
| | | (0.090) | | (0.109) |
| Temperature | | -0.065*** | | -0.066* |
| | | (0.017) | | (0.020) |
| % plain area | | 0.083*** | | 0.126* |
| | | (0.088) | | (0.101) |
| ***Census Means at Township Level*** | | | | |
| # of kids in the household | | | 0.360 | -0.021 |
| | | | (0.240) | (0.257) |
| # of youths in the household | | | -0.196 | -0.355* |
| | | | (0.132) | (0.137) |
| # of adults in the household | | | 0.427*** | 0.029 |
| | | | (0.151) | (0.163) |
| # of elderly in the household | | | 0.169 | 0.472* |
| | | | (0.256) | (0.269) |
| # HH members completed primary school | | | -0.042 | -0.005 |
| | | | (0.146) | (0.160) |
| # HH members completed junior high school | | | -0.123 | 0.068 |
| | | | (0.127) | (0.149) |
| # HH members completed senior high school | | | -0.127 | -0.287 |
| | | | (0.259) | (0.282) |
| # HH members completed vocational degree | | | 0.866 | -0.310 |
| | | | (0.642) | (0.665) |
| # HH members with college degree and above | | | 0.360 | 0.507 |
| | | | (0.562) | (0.564) |
| Housing area (meter square) | | | -0.002 | -0.003* |
| | | | (0.001) | (0.001) |
| House made of brick (dummy = 1; 0 otherwise) | | | 0.209* | 0.421* |
| | | | (0.093) | (0.118) |
| # HH members engaged in nonagricultural activities | | | -0.018** | 0.089 |
| | | | (0.097) | (0.106) |
| Married Household Head (dummy = 1; 0 otherwise) | | | 0.827* | 0.854* |
| | | | (0.458) | (0.465) |
| 3 generations living under the same roof (dummy = 1; 0 otherwise) | | | -6.206*** | -5.056* |
| | | | (2.280) | (2.447) |
| Constant | 6.703*** | 2.074** | 5.247*** | -0.910 |
| | (0.090) | (1.240) | (0.417) | (1.445) |
| Number of observations | 1070 | 1070 | 1070 | 1070 |
| R-squared | 0.211 | 0.267 | 0.247 | 0.291 |

Note: *** significant at 1%; ** significant at 5% ; * significant at 10%

**Table 3. Tests for Spatial Autocorrelation in the OLS residuals of the First Stage Consumption Model**

| Model | Moran's I | $LM_r$ | $LM_l$ |
|---|---|---|---|
| Household Characteristics | 9.895*** | 103.239*** | 82.101*** |
| Household Characteristics + Environmental Variables | 2.267* | 1.513 | 1.132 |
| Household Characteristics + Census Means | 7.365*** | -1.711 | 29.997*** |
| Household Characteristics + Environmental Variables + Census Means | 1.708* | 1.288 | 0.154 |

Note: *** significant at 1%; ** significant at 5% ; * significant at 10%

**Table 4.** First Stage Regression Model of Per Capita Expenditure (General Spatial Model)

| ***Variable*** | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| ***Household Level Characteristics*** | | | | |
| # HH members age <6 | -0.288*** | -0.331*** | -0.313*** | -0.308 |
| | (0.042) | (0.041) | (0.043) | (0.042) |
| # HH members age 7 - 15 years | -0.086*** | -0.118*** | -0.102*** | -0.103 |
| | (0.022) | (0.023) | (0.023) | (0.023) |
| # HH members age 16 - 60 years | -0.059** | -0.117*** | -0.077*** | -0.079 |
| | (0.027) | (0.026) | (0.028) | (0.028) |
| # HH members age > 60 years | -0.222*** | -0.224*** | -0.215*** | -0.229 |
| | (0.034) | (0.035) | (0.034) | (0.034) |
| # HH members completed primary school | -0.153*** | -0.095*** | -0.135 | -0.133 |
| | (0.033) | (0.033) | (0.034) | (0.033) |
| # HH members completed junior high school | -0.088** | -0.038** | -0.073 | -0.061 |
| | (0.031) | (0.031) | (0.033) | (0.032) |
| # HH members completed senior high school | -0.022 | 0.048 | -0.010 | 0.002 |
| | (0.043) | (0.045) | (0.045) | (0.045) |
| # HH members completed vocational degree | 0.148 | 0.152 | 0.178 | 0.137 |
| | (0.117) | (0.115) | (0.117) | (0.115) |
| # HH members with college degree and above | 0.255 | 0.273 | 0.225 | 0.247 |
| | (0.176) | (0.174) | (0.175) | (0.173) |
| Housing area (meter square) | 0.003*** | 0.003*** | 0.003*** | 0.003 |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| HH uses LPG as main cooking fuel (dummy = 1; 0 otherwise) | 0.301 | 0.218 | 0.352 | 0.385 |
| | (0.369) | (0.365) | (0.368) | (0.363) |
| House made of brick (dummy = 1; 0 otherwise) | 0.023 | 0.099** | -0.015 | 0.004 |
| | (0.040) | (0.044) | (0.046) | (0.047) |
| HH has access to safe drinking water (dummy = 1; 0 otherwise) | 0.145*** | 0.076* | 0.142*** | 0.087 |
| | (0.040) | (0.044) | (0.043) | (0.045) |
| HH has toilet facility in the house (dummy = 1; 0 otherwise) | 0.167*** | 0.196*** | 0.275*** | 0.271 |
| | (0.046) | (0.054) | (0.051) | (0.062) |
| # HH members engaged in non-agricultural activities | 0.141*** | 0.124*** | 0.144*** | 0.129 |
| | (0.029) | (0.029) | (0.031) | (0.031) |
| ***Environmental Variables*** | | | | |
| Total area of land | | 0.078 | | 0.165 |
| | | (0.053) | | (0.077) |
| Elevation (log) | | 0.152 | | 0.166 |
| | | (0.105) | | (0.113) |
| Density of highway (log) | | 0.030*** | | 0.040 |
| | | (0.007) | | (0.009) |
| % loam in the soil | | 0.003 | | 0.010 |
| | | (0.005) | | (0.006) |
| Annual rainfall (log) | | 0.538*** | | 0.659 |
| | | (0.126) | | (0.180) |

| | | | | |
|---|---|---|---|---|
| Slope (log) | | -0.120** | | -0.050 |
| | | (0.057) | | (0.068) |
| % organic matter in soil texture | | -0.354 | | -0.537 |
| | | (0.092) | | (0.110) |
| Temperature | | -0.045** | | -0.047 |
| | | (0.018) | | (0.022) |
| % plain area | | 0.060 | | 0.107 |
| | | (0.086) | | (0.099) |
| ***Census Means at Township Level*** | | | | |
| # of kids in the household | | | 0.291 | 0.007 |
| | | | (0.237) | (0.266) |
| # of youths in the household | | | -0.254* | -0.365 |
| | | | (0.131) | (0.137) |
| # of adults in the household | | | 0.371** | 0.085 |
| | | | (0.150) | (0.164) |
| # of elderly in the household | | | 0.184 | 0.495 |
| | | | (0.254) | (0.269) |
| # HH members completed primary school | | | 0.033 | 0.003 |
| | | | (0.143) | (0.163) |
| # HH members completed junior high school | | | -0.196 | 0.023 |
| | | | (0.126) | (0.150) |
| # HH members completed senior high school | | | -0.368 | -0.346 |
| | | | (0.253) | (0.279) |
| # HH members completed vocational degree | | | 0.177 | -0.488 |
| | | | (0.651) | (0.670) |
| # HH members with college degree and above | | | 0.016 | 0.314 |
| | | | (0.564) | (0.572) |
| Housing area (meter square) | | | -0.001 | -0.002 |
| | | | (0.001) | (0.001) |
| House made of brick (dummy = 1; 0 otherwise) | | | 0.264 | 0.452 |
| | | | (0.092) | (0.119) |
| # HH members engaged in nonagricultural activities | | | 0.114 | 0.158 |
| | | | (0.099) | (0.110) |
| Married Household Head (dummy = 1; 0 otherwise) | | | 0.558 | 0.783 |
| | | | (0.453) | (0.464) |
| 3 generations living under the same roof (dummy = 1; 0 otherwise) | | | -4.286* | -4.735 |
| | | | (2.297) | (2.474) |
| Constant | 3.140*** | 0.833 | 2.047** | -2.212 |
| | (0.600) | (1.394) | (0.740) | (1.602) |
| r | 0.597*** | 0.245** | 0.563*** | 0.250 |
| | (0.078) | (0.096) | (0.089) | (0.106) |
| l | -0.145 | -0.057 | -0.106 | -0.077 |
| | (0.119) | (0.124) | (0.086) | (0.039) |
| R-squared | 0.233 | 0.267 | 0.261 | 0.294 |

Note: *** significant at 1%; ** significant at 5% ; * significant at 10%

**Table 5.** First Stage Regression Model of Per Capita Expenditure (Spatial Error Model)

| *Variable* | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| ***Household Level Characteristics*** | | | | |
| # HH members age <6 | -0.295*** | -0.336*** | -0.309*** | -0.303*** |
| | (0.040) | (0.040) | (0.041) | (0.041) |
| # HH members age 7 - 15 years | -0.099*** | -0.1192*** | -0.1062*** | -0.1021** |
| | (0.020) | (0.023) | (0.022) | (0.022) |
| # HH members age 16 - 60 years | -0.072*** | -0.1183*** | -0.0779*** | -0.0798** |
| | (0.024) | (0.026) | (0.027) | (0.027) |
| # HH members age > 60 years | -0.221*** | -0.2262*** | -0.2212*** | -0.2299** |
| | (0.034) | (0.035) | (0.034) | (0.034) |
| # HH members completed primary school | -0.148*** | -0.098*** | -0.134*** | -0.133*** |
| | (0.032) | (0.032) | (0.033) | (0.033) |
| # HH members completed junior high school | -0.067*** | -0.033 | -0.067*** | -0.057* |
| | (0.030) | (0.031) | (0.031) | (0.031) |
| # HH members completed senior high school | 0.009*** | 0.060 | -0.001** | 0.009 |
| | (0.042) | (0.043) | (0.043) | (0.044) |
| # HH members completed vocational degree | 0.137 | 0.156 | 0.176 | 0.143 |
| | (0.115) | (0.114) | (0.113) | (0.113) |
| # HH members with college degree and above | 0.275 | 0.281* | 0.243 | 0.263 |
| | (0.172) | (0.171) | (0.171) | (0.170) |
| Housing area (meter square) | 0.003*** | 0.003*** | 0.003*** | 0.003*** |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| HH uses LPG as main cooking fuel (dummy = 1; 0 otherwise) | 0.373 | 0.261 | 0.417 | 0.413 |
| | (0.365) | (0.359) | (0.359) | (0.356) |
| House made of brick (dummy = 1; 0 otherwise) | 0.002 | 0.093** | -0.045 | -0.004 |
| | (0.039) | (0.043) | (0.045) | (0.045) |
| HH has access to safe drinking water (dummy = 1; 0 otherwise) | 0.161*** | 0.082** | 0.133*** | 0.088 |
| | (0.040) | (0.042) | (0.041) | (0.043) |
| HH has toilet facility in the house (dummy = 1; 0 otherwise) | 0.217*** | 0.207*** | 0.292 | 0.281** |
| | (0.048) | (0.054) | (0.054) | (0.061) |
| # HH members engaged in non-agricultural activities | 0.128*** | 0.111*** | 0.129 | 0.121*** |
| | (0.029) | (0.029) | (0.030) | (0.030) |
| ***Environmental Variables*** | | | | |
| Total area of land | | 0.061 | | 0.148* |
| | | (0.048) | | (0.079) |
| Elevation (log) | | 0.127 | | 0.152 |
| | | (0.100) | | (0.107) |
| Density of highway (log) | | 0.035*** | | 0.044*** |
| | | (0.007) | | (0.005) |
| % loam in the soil | | 0.001 | | 0.010* |
| | | (0.006) | | (0.006) |
| Annual rainfall (log) | | 0.681*** | | 0.785*** |
| | | (0.074) | | (0.046) |

| | | | | |
|---|---|---|---|---|
| Slope (log) | | -0.147*** | | -0.084*** |
| | | (0.055) | | (0.053) |
| % organic matter in soil texture | | -0.410*** | | -0.601 |
| | | (0.077) | | (0.094) |
| Temperature | | -0.057*** | | -0.058*** |
| | | (0.018) | | (0.020) |
| % plain area | | 0.059*** | | 0.098*** |
| | | (0.092) | | (0.101) |
| ***Census Means at Township Level*** | | | | |
| # of kids in the household | | | 0.362 | -0.001 |
| | | | (0.235) | (0.347) |
| # of youths in the household | | | -0.323** | -0.358 |
| | | | (0.130) | (0.134) |
| # of adults in the household | | | 0.312** | 0.040 |
| | | | (0.144) | (0.157) |
| # of elderly in the household | | | 0.193 | 0.463 |
| | | | (0.242) | (0.263) |
| # HH members completed primary school | | | 0.064 | 0.024 |
| | | | (0.161) | (0.159) |
| # HH members completed junior high school | | | -0.108 | 0.055 |
| | | | (0.129) | (0.147) |
| # HH members completed senior high school | | | -0.224 | -0.281 |
| | | | (0.275) | (0.282) |
| # HH members completed vocational degree | | | 0.238 | -0.327 |
| | | | (0.595) | (0.628) |
| # HH members with college degree and above | | | -0.054 | 0.496 |
| | | | (0.497) | (0.544) |
| Housing area (meter square) | | | -0.001 | -0.002 |
| | | | (0.001) | (0.001) |
| House made of brick (dummy = 1; 0 otherwise) | | | 0.252*** | 0.425 |
| | | | (0.097) | (0.112) |
| # HH members engaged in non-agricultural activities | | | 0.035 | 0.086 |
| | | | (0.097) | (0.104) |
| Married Household Head (dummy = 1; 0 otherwise) | | | 0.791** | 0.856 |
| | | | (0.336) | (0.458) |
| 3 generations living under the same roof (dummy = 1; 0 otherwise) | | | -3.584* | -4.698 |
| | | | (2.149) | (2.369) |
| Constant | 6.690*** | 2.168* | 5.497*** | -0.617 |
| | (0.025) | (1.271) | (0.071) | (1.184) |
| l | 0.479*** | 0.167*** | 0.514*** | 0.125*** |
| | (0.053) | (0.016) | (0.012) | (0.019) |
| R-squared | 0.243 | 0.269 | 0.272 | 0.292 |
| Log-likelihood function | -650.64 | -629.012 | -630.1152 | -610.5734 |

Note: *** significant at 1%; ** significant at 5% ; * significant at 10%

**Table 6.** First Stage Regression Model of Per Capita Expenditure (Spatial Lag Model)

| Variable | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| ***Household Level Characteristics*** | | | | |
| # HH members age <6 | -0.291*** | -0.338*** | -0.310*** | -0.306*** |
| | (0.042) | (0.040) | (0.038) | (0.041) |
| # HH members age 7 - 15 years | -0.092*** | -0.118*** | -0.105*** | -0.103*** |
| | (0.022) | (0.023) | (0.022) | (0.022) |
| # HH members age 16 - 60 years | -0.0633** | -0.1204*** | -0.0771*** | -0.0795** |
| | (0.027) | (0.026) | (0.022) | (0.027) |
| # HH members age > 60 years | -0.223*** | -0.224*** | -0.218*** | -0.230*** |
| | (0.034) | (0.035) | (0.028) | (0.033) |
| # HH members completed primary school | -0.158*** | -0.094*** | -0.138*** | -0.133*** |
| | (0.033) | (0.032) | (0.029) | (0.033) |
| # HH members completed junior high school | -0.076** | -0.031 | -0.066** | -0.059* |
| | (0.031) | (0.031) | (0.030) | (0.032) |
| # HH members completed senior high school | -0.007 | 0.059 | -0.006 | 0.005 |
| | (0.043) | (0.043) | (0.044) | (0.044) |
| # HH members completed vocational degree | 0.154 | 0.155 | 0.190 | 0.139 |
| | (0.116) | (0.114) | (0.098) | (0.113) |
| # HH members with college degree and above | 0.254 | 0.285* | 0.227 | 0.255 |
| | (0.175) | (0.171) | (0.172) | (0.170) |
| Housing area (meter square) | 0.003*** | 0.003*** | 0.003*** | 0.003*** |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| HH uses LPG as main cooking fuel (dummy = 1; 0 otherwise) | 0.351 | 0.229 | 0.380 | 0.396 |
| | (0.367) | (0.359) | (0.360) | (0.356) |
| House made of brick (dummy = 1; 0 otherwise) | 0.032 | 0.101 | -0.017 | 0.001 |
| | (0.041) | (0.043) | (0.043) | (0.046) |
| HH has access to safe drinking water (dummy = 1; 0 otherwise) | 0.162*** | 0.071* | 0.144*** | 0.085** |
| | (0.040) | (0.041) | (0.025) | (0.043) |
| HH has toilet facility in the house (dummy = 1; 0 otherwise) | 0.186*** | 0.212*** | 0.289*** | 0.277** |
| | (0.047) | (0.053) | (0.018) | (0.061) |
| # HH members engaged in non-agricultural activities | 0.132*** | 0.116*** | 0.137*** | 0.126*** |
| | (0.029) | (0.028) | (0.028) | (0.030) |
| ***Environmental Variables*** | | | | |
| Total area of land | | 0.062 | | 0.159** |
| | | (0.046) | | (0.076) |
| Elevation (log) | | 0.163 | | 0.166 |
| | | (0.102) | | (0.110) |
| Density of highway (log) | | 0.034*** | | 0.043*** |
| | | (0.006) | | (0.005) |
| % loam in the soil | | 0.004 | | 0.011* |
| | | (0.005) | | (0.006) |
| Annual rainfall (log) | | 0.628*** | | 0.728*** |
| | | (0.015) | | (0.021) |

| | | | | |
|---|---|---|---|---|
| Slope (log) | | -0.150** | | -0.071 |
| | | (0.050) | | (0.052) |
| % organic matter in soil texture | | -0.402*** | | -0.572*** |
| | | (0.067) | | (0.086) |
| Temperature | | -0.059*** | | -0.055*** |
| | | (0.016) | | (0.019) |
| % plain area | | 0.075 | | 0.109 |
| | | (0.086) | | (0.097) |
| ***Census Means at Township Level*** | | | | |
| # of kids in the household | | | 0.347* | 0.005 |
| | | | (0.182) | (0.243) |
| # of youths in the household | | | -0.246** | -0.364 |
| | | | (0.101) | (0.134) |
| # of adults in the household | | | 0.392*** | 0.060 |
| | | | (0.009) | (0.154) |
| # of elderly in the household | | | 0.175 | 0.483 |
| | | | (0.263) | (0.263) |
| # HH members completed primary school | | | 0.009 | 0.008 |
| | | | (0.142) | (0.157) |
| # HH members completed junior high school | | | -0.165 | 0.041 |
| | | | (0.104) | (0.147) |
| # HH members completed senior high school | | | -0.260** | -0.319 |
| | | | (0.112) | (0.276) |
| # HH members completed vocational degree | | | 0.479 | -0.407 |
| | | | (0.869) | (0.643) |
| # HH members with college degree and above | | | 0.142 | 0.409 |
| | | | (0.894) | (0.552) |
| Housing area (meter square) | | | -0.001*** | -0.002 |
| | | | (0.000) | (0.001) |
| House made of brick (dummy = 1; 0 otherwise) | | | 0.252*** | 0.441 |
| | | | (0.073) | (0.114) |
| # HH members engaged in non agricultural activities | | | 0.058 | 0.125 |
| | | | (0.058) | (0.106) |
| Married Household Head (dummy = 1; 0 otherwise) | | | 0.729 | 0.833 |
| | | | (1.700) | (0.456) |
| 3 generations living under the same roof (dummy = 1; 0 otherwise) | | | -4.973*** | -4.791 |
| | | | (0.883) | (2.405) |
| Constant | 4.510*** | 1.698 | 3.390*** | -1.481* |
| | (0.312) | (1.161) | (0.429) | (1.417) |
| r | 0.329*** | 0.070** | 0.301 | 0.132 |
| | (0.042) | (0.034) | (0.368) | (0.080) |
| R-squared | 0.205 | 0.267 | 0.245 | 0.292 |
| Log-likelihood function | -657.341 | -629.331 | -635.359 | -609.654 |

Note: *** significant at 1%; ** significant at 5% ; * significant at 10%