An Empirical Evaluation of Poverty Mapping Methodology: Explicitly Spatial versus Implicitly Spatial Approach

Susan Olivia*, John Gibson, Scott Rozelle, Jikun Huang and Xiangzheng Deng

University of California, Davis; University of Waikato; Stanford University; Chinese Academy of Sciences

Introduction

Analysts and policy makers are increasingly interested in small-area estimates of poverty. These allow governments to better target interventions to poor areas and also allow a better understanding of the determinants of poverty. But in most countries, the standard data on poverty from household surveys cannot be disaggregated below region or provincial level, while the interest is in poverty at the county or township level.

With assistance from the World Bank, analysts from at least 20 developing countries have combined detailed information from household surveys with the more extensive coverage of census data. The survey data allow models of consumption to be fitted, with the explanatory variables restricted to those available from a recent census. This method allows predicted poverty probabilities to be estimated precisely for small geographic areas (Hentschel et al., 2000).

But most applications neglect geographical and environmental components and do not explicitly account for "spatial dependencies". These dependencies arise from social interactions and unobserved factors (e.g. poor soils). Ignoring spatial autocorrelation in the regression 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 properly estimate and correct for spatial autocorrelation. Instead, these applications deal with spatial correlation by inserting local area means of census variables measuring household and individual characteristics (Elbers et al., 2003).

In this study, we use census and geo-referenced household survey data as well as environmental variables from Shaanxi, China to compare the treatment of spatial autocorrelation in poverty mapping methodology with a proper treatment based on actual location of the households and modeling the spatial dependencies with spatial regression models. Shaanxi has one of the slowest rate of poverty reduction in China and has a very heterogeneous environment.

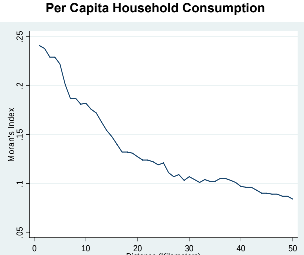
Data

We use the 2000 Population Census; the ex-post geo-referenced 2001 Household and Income Expenditure Survey that allow exact distances between each household to be measured; and a variety of spatially referenced environmental variables describing topography, soil characteristics, and land cover from satellite remote sensing data provided by the US Landsat TM/ETM images with a spatial resolution of 1 kilometer by 1 kilometer and aggregated to the county level.

Method: Spatial Regression Models

To account for spatial autocorrelation, a spatial weighting matrix W is needed to determine for each observation, which locations belong in its neighborhood. Latitude and longitude coordinates for each household in the household survey were used to calculate this and then the Moran's Index of spatial autocorrelation in household per capita consumption was calculated for varying neighborhood sizes of 1 – 50km (Figure 1).

Figure 1. Spatial Correlation in Per Capita Household Consumption



When households within a one kilometer radius of location i are considered as the neighborhood, Moran's I = 0.24 and is statistically significant ($p<0.01$). In other words, a regression of spatially weighted average per capita consumption within this neighborhood, W_s on per capita consumption of each household s, would have a statistically significant coefficient of 0.24. For all neighborhood sizes considered, Moran's I is still statistically significant.

Given the evidence of spatial autocorrelation in the data, we then estimate a model of per capita household consumption using:

1. Spatial Lag Model: $Y = \rho WY + X\beta + \varepsilon$
where X is the vector of explanatory variables for the household and WY is the spatially lagged dependent variable.

Unlike time series lagged dependent variable models, *OLS is always inconsistent*, because $(WY)_i$ correlated with both ε_i and errors in other locations.

2. Spatial Error Model: $Y = X\beta + \varepsilon,$ where $\varepsilon = \lambda W\varepsilon + \mu$

In this model, error for one observation depends on weighted average of the errors for neighboring observations, as a result, OLS estimates are no longer efficient.

The household-level variables used in each model are:

- type and size of dwelling
- household size and age composition
- proportion of adults in the household with primary, secondary and higher education
- indicator for households who engaged in non-agricultural sector
- indicator for households who use LPG as main cooking fuel

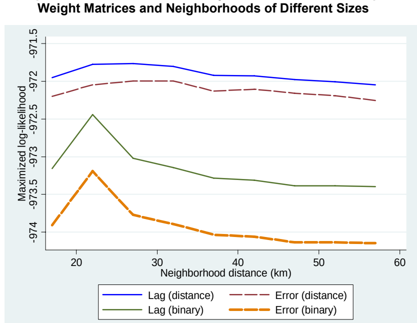
All of these variables are also available in the census and so township level means of these same variables are included. These means help to drive cluster specific smaller and improve precision of the second stage of predictions (Elbers et al., 2003).

We also augment our models with environmental variables from satellite remote sensing data. The environmental variables used in our study are percentage of arable land in the county, annual rainfall, temperature, elevation, slope and soil characteristics.

Spatial Regression Results

In light of the above results about the misspecification when OLS is used, a variety of spatial lag and spatial error models of per capita consumption were estimated for 1058 households using data from the household survey. Nine neighborhood sizes were chosen, ranging from the minimum feasible (to prevent 'islands' with no neighbors) of 17 kilometers to a maximum of 57 kilometers. Both binary and inverse distance weights were considered.

Figure 2. Log-Likehood Values for Spatial Lag and Spatial Error Models with Binary and Inverse Distance Spatial Weight Matrices and Neighborhoods of Different Sizes



A comparison of the maximized log likelihoods of the resulting models indicated that there was better performance for the spatial lag model than the spatial error model (Figure 2). The log likelihood is higher when the spatial weight matrix was based on inverse distance rather than a simple 0/1 weights. Grid search favors estimating spatial lag model of 27 kilometers.

Table 1 shows the results of regressing per capita consumption on three types of unrestricted exogenous variables: household characteristics, census means of household characteristics and environmental variables from the spatial error and spatial lag models (based on inverse distance weights and a neighborhood size of 27 kilometers).

According to the maximum likelihood estimates, in the preferred spatial lag specification, the spatial parameter corresponding to the spatial lag model (ρ) is statistically significant across 3 sets of estimation. Using only the household characteristics, $\rho = 0.32$ indicates that on average, a 10 percentage point increase in per capita consumption in a particular location will result in a 3.2 percentage point increase in the per capita consumption in a neighboring location.

The results reported in Table 1 also allows us to assess whether census means and/or environmental variables can be used to soak up the unwanted spatial autocorrelation in poverty mapping methodology. There seems evidence that including census means in the first stage model of consumption reduces the spatial autocorrelation by 12 percent. However, if we augmenting the model with both census means and environmental variables, it further reduces the unwanted spatial autocorrelation by almost 36 percent.

Despite the fact that integrating both census means and environmental variables prove to be effective in reducing the spatial autocorrelation in the data, our results suggests that these variables did not soak up all the spatial autocorrelation, implying that there is a need to explicitly model the spatial autocorrelation, if not OLS model is likely to be mis-specified.

Table 1. Spatial Error and Spatial Lag Estimates of Per Capita Household Consumption

	HH Vars	HH Vars+ Census Means	HH Vars + Env Vars + Census Means
ρ (spatial lag)	0.322***	0.283***	0.207***
λ (spatial error)	0.346***	0.298***	0.215***

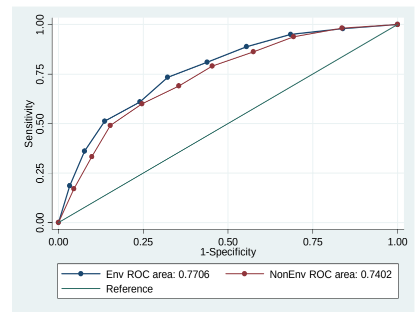
Our results also suggest that consumption is higher for households with larger dwellings, with a greater proportion of adults who completed senior high school education and above, engaged in the non-agricultural sector and when using LPG as main cooking fuel. Consumption per capita is lower for larger households. 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.

Inclusion of environmental variables raises the R^2 of the consumption model from 0.23 to 0.29 and these variables are jointly statistically significant with a F-value of 7.81, suggesting that consumption is highly related to the characteristics of the environment where people live. The environmental variables show that consumption is lower for households living in areas that are on steep slopes. Furthermore, households residing in areas covered with lower percentage of loam and with lower rainfall tend to have lower consumption than those residing in other areas.

Assessing Performance of Poverty Map with Environmental Variables

Given knowledge about where poor people or households live, a natural question to ask is how effective environmental variables are in identifying the poor. To assess if targeting is likely to be more accurate when environmental variables are used to construct the poverty map, we plot the Receiver Operating Characteristic (ROC) curves. A ROC curve plots the probability of a variable correctly classifying a poor household as poor on the vertical axis ('the sensitivity') against one minus the probability of the same variable correctly classifying a non-poor household as non-poor household (1-specificity) on the horizontal axis. The closer a ROC curve is to the 45° line, the weaker is the diagnostic accuracy of the variable that is being considered as a targeting indicator. Figure 3 shows that the ROC curve without the environmental variables is significantly ($p<0.001$) closer to the 45° line.

Figure 3. Comparing the Targeting Performance of Consumption Models With and Without Environmental Variables



Conclusions

We take an explicit spatial econometric approach in estimating specifications that incorporate spatial dependency in the first stage of consumption model of poverty mapping exercises. The significance of the spatial parameters indicates that these dependencies should be incorporated in the poverty mapping methodology because ignoring spatial error structure can cause inference problems, while ignoring spatial lags can bias coefficient estimates. Our results show that both census means and environmental variables can go a considerable way towards removing spatial autocorrelation, however, these control variables did not soak up all of the unwanted spatial autocorrelation. The results suggest that in order to explicitly model the spatial effects, analysts need to know actual distance between households, which in this sense are supportive of the usage of GPS in household surveys.

Notes

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

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 – 165.

* Corresponding author: solivia@ucdavis.edu. More information on this and related projects can be obtained at www.chinapovmap.org.

Acknowledgements

We are grateful to Kiwi Advanced Research and Education Network for Financial support and to Albert Park, Hongbin Li and Shuming Bao for assistance.