UNIVERSITY OF WAIKATO

Hamilton
New Zealand

Can We Trust Cluster-Corrected Standard Errors?
An Application of Spatial Autocorrelation
with Exact Locations Known

John Gibson
University of Waikato

Bonggeun Kim
Seoul National University
and

Susan Olivia
Monash University

Department of Economics

Working Paper in Economics 10/07

November 2010

Corresponding Author

John Gibson
Department of Economics
University of Waikato,
Private Bag 3105,
Hamilton, New Zealand

Fax: +64 (7) 838 4331
Email: jkgibson@waikato.ac.nz

Abstract

Standard error corrections for clustered samples impose untested restrictions on spatial correlations. Our example shows these are too conservative, compared with a spatial error model that exploits information on exact locations of observations, causing inference errors when cluster corrections are used.

Keywords

clustered samples
GPS
spatial correlation

JEL Classification

C31, C81

Acknowledgements

We are grateful to Yun Liang for assistance with the computer programming used in this paper. All remaining errors are our responsibility.

1. Introduction

Inference methods that recognize the clustering of individual observations are now widely used in applied econometrics (Wooldridge, 2003). An early, cautionary, example of distorted inferences when ignoring the potential correlation between observations sharing the same cluster was provided by Pepper (2002). Yet with continuing changes in the technology of survey data collection, it is possible that clustered standard errors are now too widely used, causing a new set of distorted inferences.

Increasingly, household surveys geo-reference exact locations (within 15 meter accuracy) of respondents, using the Global Positioning System (GPS). This is especially in developing countries, where face-to-face surveying predominates so dwellings are easily geo-referenced when interviewers visit households, and where the falling cost and improved accuracy of GPS receivers has most increased demand for location data (Gibson and McKenzie, 2007). In this note, we question whether the usual inference methods for dealing with clustered samples remain the best option when econometricians know exact locations, rather than just that groups of observations share the same cluster.

We first use a simple spatial error model to show the untested restrictions that clustered standard errors place on spatial correlations. We then provide an example from a geo-referenced household survey in Indonesia where inferences about village-level determinants of income from non-farm rural enterprises (NFRE) are distorted by using clustered standard errors. These NFRE are an important escape path from rural poverty and are heavily affected by location-specific investments in infrastructure and the quality of the business environment (Isgut, 2004). Hence, correct inferences about drivers of NFRE activity can be very useful to economists and policy makers interested in rural poverty.

2. Robust Standard Errors for Clusters and Spatial Correlation

To show the restrictions on spatial correlations from the typical (robust) cluster correction, we consider a simple model with a string city, equal distance between respondents, and first-order (positive) spatial correlations (λ, ρ) of errors in a simple linear regression, $y = \beta_0 + \beta_1 X + u.$ The variance of the slope coefficient $\hat{\beta}_1$ of the population regression is:

$$V(\hat{\beta}_1)=\frac{V(\bar{u})}{(\sigma_x^2)^2},\ V(\bar{u})=\frac{\sigma_u^2}{N}[1+2N_c\sum_{j=1}^{m-1}\frac{m-j}{N}\lambda^j]+\frac{\sigma_u^2}{N}\{[2(N_c-1)\sum_{j=1}^{m-1}\frac{j}{N}\rho^j]+[2\sum_{j=m+1}^{N-1}\frac{N-j}{N}\rho^j]\}\ (1)$$

where N_c is the total number of clusters, m is the total number of observations, *j* in a cluster, and $N=m\times N_c$. The first term in $V(\bar{u})$ is the sum of the covariances within a cluster, with intra-cluster spatial correlation, λ. The second term involves the inter-cluster correlation, $\rho(\leq\lambda)$.

Cluster corrections make no allowance for spatial correlations between observations in different clusters, imposing the untested restriction $\rho=0$. But in reality, such correlations may not vanish, as recently shown for the example of State-level variables in the U.S. (Barrios et al., 2010). Moreover, since spatial correlations within clusters are rarely known, cluster corrections assume the same intra-cluster correlation between any two error terms, $corr(u_{jc},u_{j'c})=\gamma$ for $j\neq j'$. But rural clusters are often of quite unequal area and the strength of common factors shared by observations in the same cluster may vary with environmental and economic heterogeneity. Hence intra-cluster correlations in errors may vary with population density and the strength of omitted common factors.

With these restrictions imposed, equation (1) becomes the clustered estimator:

$$V(\hat{\beta}_{1,C})=\frac{V(\bar{u}_C)}{(\sigma_x^2)^2},\ V(\bar{u}_C)=\frac{\sigma_u^2}{N}[1+2N_c\sum_{j=1}^{m-1}\frac{m-j}{N}\hat{\gamma}] \qquad (2)$$

Note that $\text{var}(\hat{\beta}_{1,C})\gtrless\text{var}(\hat{\beta}_1)$ since:

$$\frac{1}{(\sigma_x^2)^2}\frac{\sigma_u^2}{N}[1+2N_c\sum_{j=1}^{m-1}\frac{m-j}{N}(\hat{\gamma}-\lambda^j)]\gtrless\frac{1}{(\sigma_x^2)^2}\frac{\sigma_u^2}{N}\{[(N_c-1)\sum_{j=1}^{m-1}\frac{j}{N}\rho^j]+[\sum_{j=m+1}^{N-1}\frac{N-j}{N}\rho^j]\}. \quad (3)$$

When the right-hand side of equation (3) is negligible, as with $\rho\to 0$, we expect $\text{var}(\hat{\beta}_{1,C})>\text{var}(\hat{\beta}_1)$ from the efficiency gain when using the precise weighted least squares error terms for first-order spatial correlations, rather than assuming the same spatial correlation within every cluster. Note also that $V(\hat{\beta}_{1,C})>V(\hat{\beta}_1)>V(\hat{\beta}_{1,OLS})$ where $V(\hat{\beta}_{1,OLS})$ is the case where potential correlations between disturbances (whether in the same cluster or not) are ignored.

3. Application

To investigate effects of the restrictions imposed by the standard cluster correction, we use clustered data from a geo-referenced household survey in Indonesia to estimate an income share equation for net earnings from non-farm rural enterprises. The key features of the Rural Investment Climate Survey (RICS) are clustering, with our sample of 1600 rural households located in 97 clusters, and geo-referencing of every household by GPS. The survey was fielded in only six of Indonesia's 370 districts (*kabupaten*) so clusters within each district are closer together than for a similarly sized national survey. The survey includes both household-level and community-level variables; since community variables are common to all households in a cluster, inferences about them may be especially susceptible to misspecification of the spatial correlations between errors.

To illustrate the spatial scale of correlations in the income shares we estimate Moran's I

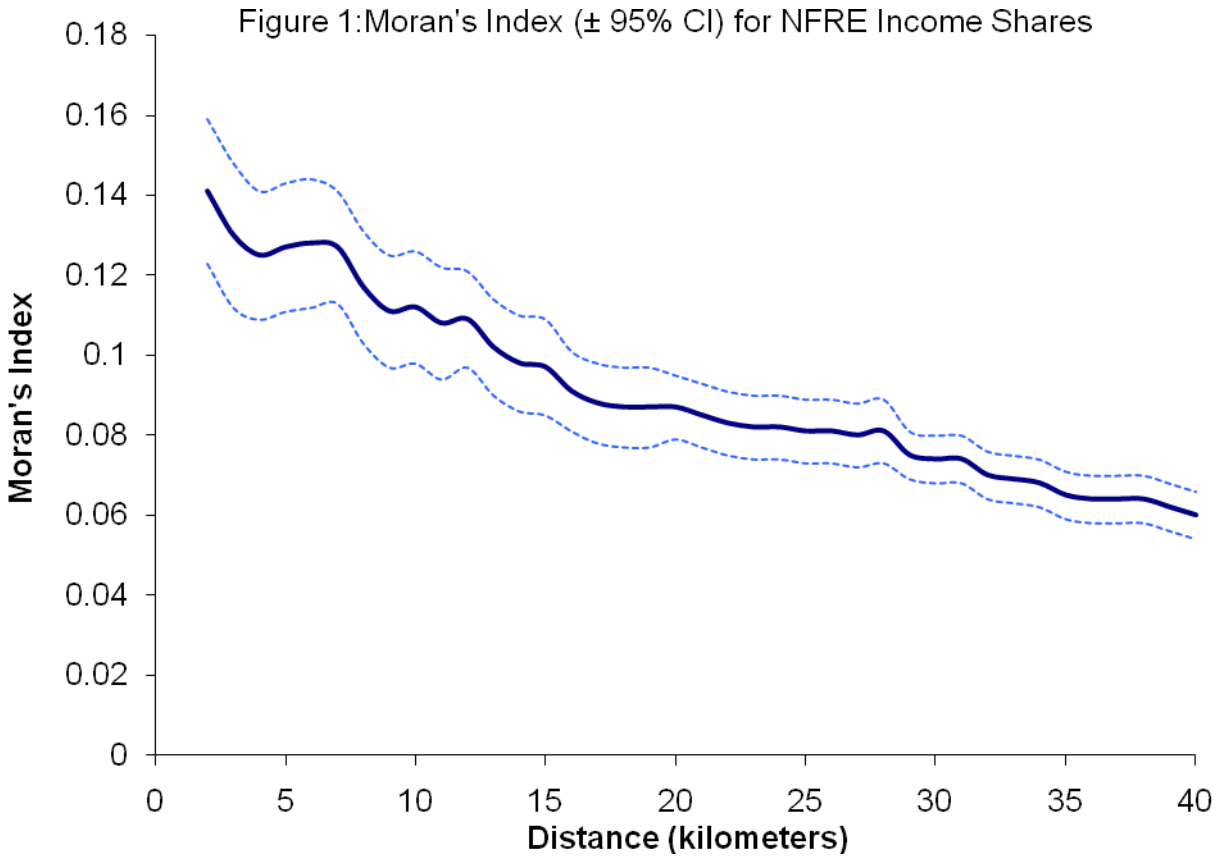
$$I = \frac{\mathbf{y}'\mathbf{W}\mathbf{y}}{\mathbf{y}'\mathbf{y}} \tag{4}$$

where **y** is a vector of income shares, **W** is the (row-standardized) spatial weight matrix, with w_{ij}=0 for non-neighbours and otherwise $w_{ij} = 1/d_{ij}$ where d_{ij} is the distance between observations i and j (inverse distance weights). This is equivalent to a regression of the spatially weighted average of income shares within a neighbourhood on the income share for each household.

Latitude and longitude coordinates were used to calculate d_{ij} for every household, for varying neighbourhood sizes of 1-40 km. The average distance from each household to the cluster center is only 0.8 km and the largest distance between any two households in a given cluster averages 1.9 km. Hence this range allows for correlations that extend far beyond the boundary of clusters. For all neighbourhood sizes considered, Moran's I is statistically significant, ranging from 0.15 at 1 km to 0.09 at 20 km and 0.06 at 40 km (Figure 1).

To see if spatial correlations extending beyond cluster boundaries are also apparent in OLS residuals, an income share model was estimated with explanatory variables typically used in the NFRE literature. These included attributes of the household head (age, gender, religion, marital status, education), and the household (size, composition, land ownership,

income), and community characteristics. The community variables are of most interest; these are common to all households in a cluster so inferences about them may be sensitive to mis-specified spatial correlations between errors. Moreover, factors such as village infrastructure and quality of the business environment may be more amenable to intervention than are individual characteristics, giving policy salience to these community variables.



The OLS results suggest that households in larger villages with a business association have higher NFRE income shares. In villages further from both cooperatives and the sub-district headquarters, experiencing crime or other disputes, and with a low occurrence of electricity blackouts, households have lower NFRE income shares (Table 1, column 1). But, while standard errors from this OLS model are heteroscedastically-robust, they ignore potential correlations between disturbances (whether in the same cluster or not), and so may be misleading.

In fact when Moran's *I* is estimated for these OLS residuals, there is always a statistically significant (p<0.01) spatial correlation, for neighbourhoods extending from 1 km to 40 km.[1] Hence, the spatial correlation in the dependent variable shown in Figure 1 is not removed by the covariates, making the inferences from the OLS results potentially misleading, even with robust standard errors. The spatial scale considered extends well beyond cluster boundaries, implying that the restriction imposed by the usual correction for clustering, of $\rho = 0$, does not hold.

Table 1: OLS, Clustered, and Spatial Error Estimates

Community Variables	Robust std errors	Clustered std errors	Spatial error model	Spatial error (ρ=0)
log(# of households in village)	0.102	0.102	0.097	0.101
	(0.0231)**	(0.0442)*	(0.0299)**	(0.0302)**
Village has business association	0.103	0.103	0.117	0.112
	(0.0299)**	(0.0636)	(0.0391)**	(0.0385)**
Village had crime/dispute last year	-0.080	-0.080	-0.078	-0.080
	(0.0224)**	(0.0314)*	(0.0299)**	(0.0301)**
Village has a cooperative	0.040	0.040	0.039	0.042
	(0.0245)	(0.0374)	(0.0328)	(0.0323)
Distance to cooperative (km)	-0.490	-0.490	-0.451	-0.466
	(0.1788)**	(0.2344)*	(0.2464)+	(0.2415)+
Distance to sub-district (km)	-1.284	-1.284	-1.314	-1.342
	(0.6688)+	(0.9727)	(0.9062)	(0.8947)
Low blackouts (< 30 minutes/day)	-0.051	-0.051	-0.054	-0.053
	(0.0282)+	(0.0481)	(0.0372)	(0.0366)
Village has no telephones	0.057	0.057	0.056	0.059
	(0.0404)	(0.0633)	(0.0544)	(0.0532)
Village has unsealed roads	0.041	0.041	0.044	0.045
	(0.0293)	(0.0446)	(0.0386)	(0.0388)
Phi (spatial autoregressive parameter)			0.285	0.271
			(0.046)**	(0.0384)**
R-squared	0.16	0.16		
Log-likelihood function	-566.32	-566.32	-541.34	-541.68

Notes: Standard errors in ( ). **=p<0.01, *=p<0.05, +=p<0.10. Characteristics of the household head (age, gender, religion, marital status, education) and the household (size, composition, land ownership, income) also included.

When the clustered standard errors are calculated (Table 1, column 2) they exceed the robust standard errors, by 47 percent on average. Moreover, three community variables (having a business association, distance to sub-district headquarters and blackouts) appearing statistically significant with the robust standard errors now appear insignificant.

[1] The evidence of statistically significant spatial autocorrelation in the OLS residuals is also apparent from Lagrange Multiplier tests, for all neighbourhood sized considered. Results of these tests are available from the authors.

The first two columns of Table 1 ignored the GPS information on exact locations. To exploit this extra information we estimate a spatial error model:

$$\begin{aligned} Y &= X\beta + u \\ u &= \varphi W u + \varepsilon \end{aligned} \tag{5}$$

where φ is the spatial autoregressive coefficient, ε a vector of iid errors and everything else is as defined above. In this model, the error for one observation depends on a weighted average of the errors for neighbouring observations (irrespective of whether in the same cluster or not). After experimenting with neighbourhoods of different sizes, a 10 km neighbourhood was found to maximize the log-likelihood and resulted in a spatial autoregressive estimate of φ=0.29 (Table 1, column 3). In other words, the spatially weighted residual NFRE share within a 10 km radius is significantly associated with the residual income share for a particular household even after controlling for household characteristics and a set of location attributes.

When the spatial error model is used, standard errors are mostly smaller (by 21 percent, on average) than for the clustered standard errors. Moreover, one of the indicators of the quality of the local business environment, whether there is a village business association, has a strongly significant (p<0.01) effect on income from non-farm rural enterprises. Yet when the cluster correction was used, the standard error on the business association indicator was almost twice as large and it appeared as an insignificant determinant of NFRE income shares.

The standard cluster correction imposes two restrictions; that inter-cluster correlations vanish (ρ=0), and that intra-cluster correlations are the same everywhere irrespective of cluster area, density of observations and importance of shared unobservable factors for neighbours. To see which of these two sources is more important to the smaller standard errors and changed inferences when moving from the cluster correction to the spatial error model, we estimate a spatial error model where all weights are set to zero for pairs of observations not in the same cluster.

The results in the last column of Table 1 that rely on the restriction that ρ=0 are almost identical to the results in column 3 where no restrictions were placed on the spatial error model. This comparison suggests that most of the overstatement of standard errors when

using the standard cluster correction comes from assuming the wrong form of spatial correlation within clusters, rather than from the implicit assumption that inter-cluster correlations vanish.

4. Conclusions

The widely used standard error correction for clustered surveys imposes untested restrictions on spatial correlations. The resulting clustered standard errors are too conservative, compared with those coming from a spatial error model that uses exact locations of observations. In our example, the main source of error was from assuming the wrong form of spatial correlation within clusters, rather than from the implicit assumption that inter-cluster correlations vanish. These results suggest that more robust inferences are likely to come from knowing actual distance between observations, supporting the growing use of GPS in household surveys to identify neighbours and the strength of their interactions.

References

Barrios, T, Diamond, R., Imbens G., and Kolesar, M. (2010). "Clustering, spatial correlations and randomization inference" *NBER Working Paper* No. 15760.

Gibson, J and McKenzie, D. (2007) "Using the Global Positioning System (GPS) in household surveys for better economics and better policy" *World Bank Research Observer* 22(2): 217-241.

Isgut, A. (2004). "Non-farm income and employment in rural Honduras: assessing the role of locational factors" *Journal of Development Studies* 40(3): 59-86.

Pepper, J. (2002). "Robust inferences from random clustered samples: An application using data from the Panel Study of Income Dynamics" *Economics Letters* 75(3): 341-345.

Wooldridge, J. (2003). "Cluster-sample methods in applied econometrics" *American Economic Review* 93(2), 133-138.