Economics Bulletin

Volume 34, Issue 3

Cluster-Corrected Standard Errors with Exact Locations Known: An Example from Rural Indonesia

John Gibson
University of Waikato

Bonggeun Kim
Seoul National University

Susan Olivia
Monash University

Abstract

Cluster-corrected standard errors are widely used but may sometimes be inappropriate since household surveys are increasingly geo-referenced. Compared with the appropriate spatial error models that use details on exact locations, cluster corrections impose untested restrictions on spatial correlations. Our example from rural Indonesia suggests cluster corrections are too conservative and may cause inference errors.

We are grateful to participants at the Spatial Econometrics Association meetings, two anonymous referees and the editor for helpful comments. All remaining errors are our own.

Citation: John Gibson and Bonggeun Kim and Susan Olivia, (2014) "Cluster-Corrected Standard Errors with Exact Locations Known: An Example from Rural Indonesia", *Economics Bulletin*, Vol. 34 No. 3 pp. 1857-1863.

Contact: John Gibson - jkgibson@waikato.ac.nz, Bonggeun Kim - bgkim07@snu.ac.kr, Susan Olivia - susan.olivia@monash.edu.

Submitted: January 30, 2014. **Published:** August 20, 2014.

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 changes in the technology of survey data collection mean that use of clustered standard errors may sometimes be inappropriate and cause inference errors. Increasingly, household surveys geo-reference 'exact' (that is, within 15 meter accuracy) locations 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). Moreover, it is in developing countries where spatial clustering of economic outcomes is most pronounced because of the importance of environmental heterogeneity to livelihoods.

In this paper, we question whether the usual inference methods for dealing with clustered samples remain the best option when practitioners 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

We consider an economic model where the unobservable behavior of neighboring households affects own-behavior through social proximity. Previous examples of such models include competitive spatial pricing in real estate (Haining, 1984) and spatially correlated household demand due to interdependent preferences (Case, 1991). In both examples, and more generally, spatial proximity is used as a proxy for unobservable social proximity. The spread of non-farm rural enterprises also likely depends on such proximity, due either to learning from neighbors or to the need for coordination when beginning a new activity with uncertain market prospects.

In general, the unknown spatial correlation patterns in such models are assumed to decay with distance. But they also could follow a more clustered pattern, with all close neighbors having similar correlations and more distant neighbors having zero correlation. This is the pattern assumed by the standard approach to estimating clustered standard errors. To more formally illustrate the restrictions on spatial correlations that such clustering entails, we consider a simple model with households located along a line (such as a road), equal distance between respondents, and first-order spatial correlations (λ, ρ) of errors in a simple linear regression, $y = \beta_0 + \beta_1 X_j + u_j$. To do so, we compare standard errors of the regression slope coefficient estimator under the first-order spatial correlation with the corresponding estimator for clustered standard errors. The variance estimator that is consistent to spatial correlation is:

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

where *j* indexes households, N_c is the total number of clusters, m is the number of observations 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, $\rho(\leq \lambda)$.[1]

Cluster corrections make no allowance for spatial correlations between different clusters, imposing the untested restriction $\rho = 0$. But in reality, such correlations may not vanish, as noted by Elbers et al. (2008). Moreover, since spatial correlations within clusters are often unknown, cluster corrections assume the same intra-cluster correlation between any two error terms, $corr(u_{jc}, u_{j'c}) = \gamma$ for $j \neq j'$. Yet in practice, survey clusters in rural places can be quite unequal in area and may exhibit considerable environmental and economic heterogeneity. Hence intra-cluster correlations in errors may vary with cluster size and population density and with the strength of omitted common factors, rather than being constant as standard cluster correction methods assume.

Imposing the restrictions that ρ=0 and $corr(u_{jc}, u_{j'c}) = \gamma$ for $j \neq j'$, equation (1) becomes the widely used cluster-corrected variance estimator:

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

Note that $V_c(\hat{\beta}_1) \underset{<}{>} V(\hat{\beta}_1)$ if:

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

When the right-hand side of equation (3) is negligible, as with $\rho \to 0$, we expect that $V_c(\hat{\beta}_1) > V(\hat{\beta}_1)$ due to the efficiency gain that comes from using the more precise weighted least squares estimator for models with first-order spatial correlations, rather than making the assumption that there is the same spatial correlation pattern within every cluster.

3. An Example from Rural Indonesia

To investigate effects of the restrictions imposed by the standard cluster-corrected variance estimator, 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.

[1] Equation (1) is a modified formula for heteroskedasticity and autocorrelation consistent (HAC) standard errors (Stock and Watson, 2007, p.606). Although the standard errors of equation (1) are spatial correlation consistent only, they can also be considered as heteroskedasticity and spatial correlation consistent standard errors if the error term (u_j) in the equation is replaced by $v_j (= (X_j - \mu_X)u_j)$.

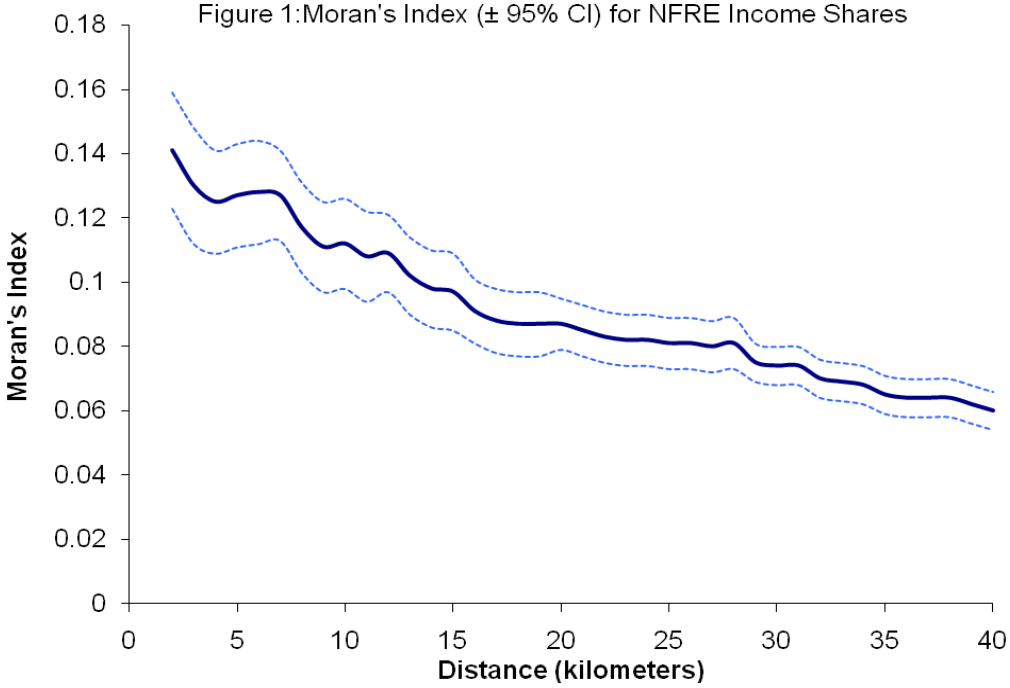
As noted above, the phenomena that we study – the share of household income from non-farm rural enterprises – is potentially affected by both the spatial and social proximity of neighbors. Social proximity may matter because of learning effects. Whether the spatial scale of learning coincides with the scale of survey clusters is an empirical question. Spatial proximity may matter because of local political economy – corruption is often considered a barrier to business development in Indonesia – and survey clusters do not always overlap local political boundaries. Finally, environmental heterogeneity may affect non-farm enterprises over spatial scales that differ from the scale of clustering and with a varying importance across clusters.

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 $\mathbf{y}$ is a vector of income shares, $\mathbf{W}$ is the (row-standardized) spatial weight matrix, with w_{ij}=0 for non-neighbors 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 neighborhood on the income share for each household. Latitude and longitude coordinates were used to calculate d_{ij} for every household, for varying neighborhood sizes of 1-40 km. The average distance from each household to the cluster centre 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 neighborhood 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).

Figure 1:Moran's Index (± 95% CI) for NFRE Income Shares



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 cooperatives and from the sub-district headquarters, experiencing crime or other disputes, households have lower NFRE income shares (Table 1, column (1)). But, while the reported standard errors from this OLS model are heteroskedasticity-robust, they ignore potential correlations between disturbances (whether in the same cluster or not), and so may be misleading.

Table 1: Regression Estimates, With Standard Errors from Robust, Clustered and Spatial Error Estimators

Community Variables	OLS, robust std errors (1)	Clustered std errors (2)	Spatial error model (3)	Spatial error (ρ=0) (4)
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 ( ). The standard errors for OLS in column (1) are heteroskedasticity-robust but otherwise ignore clustering and spatial locations. **=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.*

In fact when Moran's I is estimated for these OLS residuals, there is always a statistically significant (p<0.01) spatial correlation, for neighborhoods extending from 1 km to 40 km.[2] In

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

other words, the spatial correlation in the dependent variable that is shown in Figure 1 is not removed by the covariates, making the inferences from the OLS results potentially misleading, even with heteroskedasticity-robust standard errors. Moreover, the spatial scale considered with Moran's I extends well beyond cluster boundaries, implying that the restriction imposed by the usual correction for clustering, of $\rho = 0$, also does not hold.

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

The results in column (1) and (2) 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 neighboring observations (irrespective of whether in the same cluster or not). After experimenting with neighborhoods of different sizes, a 10 km neighborhood 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 average residual income 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 smaller than when the cluster-corrected variance estimator is used for the OLS regression coefficients, for all covariates but one (distance to the cooperative, where standard errors are higher by five percent with the spatial error model). On average, over all the covariates in Table 1, standard errors are 21 percent smaller with the spatial error model than with the clustered errors. Moreover, one of the indicators of the quality of the local business environment, whether there is a village business association, appears to have a strongly significant (p<0.01) effect on income from non-farm rural enterprises when standard errors are either heteroskedasticity-robust or from the spatial error model but when the cluster correction was used, the standard error on the business association indicator was almost twice as large and it appeared as a statistically 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 neighbors. To see which of these two restrictions 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. In other words, it appears that intra-cluster correlations are not the same everywhere, and that instead

they may vary with factors such as the size of clusters, the density of observations and the importance of proximity for sharing unobservable factors between neighbors.

4. Conclusions

The widely used cluster-corrected variance estimator imposes untested restrictions on the pattern of spatial correlations. In our example, the resulting clustered standard errors are too conservative, compared with those coming from a spatial error model that uses exact locations of observations. On average, standard errors were smaller by one-fifth when the spatial error model that utilizes the location-specific coordinates was used. The larger standard errors when the cluster correction is applied would also cause an inference error, with one of the covariates that is most amenable to being changed by policy interventions (having a village business association) appearing to be statistically insignificant in the cluster-corrected results, but highly significant when all of the other variance estimators were used.

The main source of overstatement in the clustered standard errors 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 neighbors, rather than just accounting for the fact that groups of observations share the same cluster.

References

Case, A. (1991) "Spatial patterns in household demand" *Econometrica* **59**, 953-965.

Elbers, C., Lanjouw, P., and Leite, 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 the Global Positioning System (GPS) in household surveys for better economics and better policy" *World Bank Research Observer* **22**, 217-241.

Haining, R. (1984) "Testing a spatial interacting-markets hypothesis" *Review of Economics and Statistics* **66**, 576-583.

Isgut, A. (2004) "Non-farm income and employment in rural Honduras: assessing the role of locational factors" *Journal of Development Studies* **40**, 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**, 341-345.

Stock, J. and Watson, M. (2007) *Introduction to Econometrics* 2nd ed. Pearson Education, Inc.

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