Is multiscalar analysis a panacea?

The impacts of the MAUP on omission error

YE, Xiang; ROGERSON, Peter A.

Department of Geography, University at Buffalo, the State University of New York

1. When multiscalar analysis fails

The modifiable areal unit problem (MAUP) is one of the most long-standing and far-flung problems in geography, indicating the sensitivity or inconsistency (Openshaw and Taylor 1981; Openshaw 1983; Fotheringham and Wong 1991; Larsen 2000) of the results of spatial analyses in the same study area with different spatial configurations.

When a spatial analysis is impeded by the MAUP, one of the solutions is to report the analysis results on multiple scales to mitigate the impacts (Hennerdal and Nielsen 2017). But this may not be always a good solution for research that based on regression analysis: if one or more independent variables are missing from the regression model, i.e. there is an omission error, the bias of the coefficient estimates due to the omission error will be distorted differently on different scales by different spatial configurations. In that case, multiscalar analysis dose not help; every estimate is a biased estimate, and the bias is not guaranteed to be monotonic.

2. The omission error at the individual and aggregated levels

When there is no MAUP, it is known that the expectation of the coefficient estimator is biased with the presence of an omission error (Greene 2012, p. 96, Equation 4-10):

$$\mathbb{E}(\tilde{\boldsymbol{b}}|\boldsymbol{X}) = \boldsymbol{\beta}_1 + \tilde{\boldsymbol{F}}\boldsymbol{\beta}_2 \tag{1}$$

Here, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are the true coefficients for the remaining and missing independent variables, while $\tilde{\boldsymbol{b}}$ is the corresponding (incorrect) estimator for $\boldsymbol{\beta}_1$. $\tilde{\boldsymbol{F}}$ is a $k_1 \times k_2$ matrix, characterizing the multicollinearity between the remaining independent variables $\boldsymbol{X}_1$ and missing independent variables $\boldsymbol{X}_2$ (Greene 2012, p. 96, Equation 4-11):

$$\tilde{\boldsymbol{F}} = \left(\boldsymbol{X}_1{}^{\mathrm{T}}\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1{}^{\mathrm{T}}\boldsymbol{X}_2 \tag{2}$$

Equation (2) suggests that, when independent variables $\boldsymbol{X}_2$ is dropped from the model, its effect on $\boldsymbol{y}$, namely $\boldsymbol{\beta}_2$, still exists and will be partially transferred to the (incorrectly) estimated coefficient $\tilde{\boldsymbol{b}}$, at the intensity of $\tilde{\boldsymbol{F}}$. This typically makes $\tilde{\boldsymbol{b}}$ a biased estimator for $\boldsymbol{\beta}_1$ and it is the major consequence of the omission error.

Both the MAUP and the omission error can happen simultaneously. When the MAUP is presenting, the expectation of the estimator of $\boldsymbol{\beta}_1$ at the aggregated level, $\tilde{\boldsymbol{b}}^{\circ}$, will be

$$\mathbb{E}(\tilde{\boldsymbol{b}}^{\circ}|\boldsymbol{X},\boldsymbol{M}) = \boldsymbol{\beta}_{\mathbf{1}} + \tilde{\boldsymbol{F}}\boldsymbol{\beta}_{\mathbf{2}} + \tilde{\boldsymbol{H}}^{\circ}\boldsymbol{\beta}_{\mathbf{2}} \tag{3}$$

Here, $\tilde{\boldsymbol{H}}^{\circ}$ is a $k_1 \times k_2$ matrix revealing the aggregated-level multicollinearity between the remaining and missing independent variables, after their individual-level multicollinearity, $\tilde{\boldsymbol{F}}$, has been controlled:

$$\tilde{\boldsymbol{H}}^{\circ} = \left(\boldsymbol{X}_1^{\circ}{}^{\mathrm{T}}\boldsymbol{X}_1^{\circ}\right)^{-1}\boldsymbol{X}_1^{\circ}{}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{U} \tag{4}$$

In equation (4), $\boldsymbol{X}_1^{\circ}$ is an $m \times k_1$ matrix, representing the remaining independent variables observed at the aggregated level. $\boldsymbol{M}$ is an $m \times n$ matrix, describing the effect of the MAUP, when n individual observations in the study area are merged into m regions (details omitted here). $\boldsymbol{U}$ is an $n \times k_2$ matrix, formed by combining all the residual vectors $\boldsymbol{u}_j$ at the individual level; and $\boldsymbol{u}_j$'s are the residuals generated by regressing each column of $\boldsymbol{X}_2$ on the entire $\boldsymbol{X}_1$.

3. Understanding the impacts of the MAUP on the omission error

By probing equation (3), the following understandings regarding the omission error at the aggregated level can be drawn.

First, the expectation of coefficient estimations, $\mathbb{E}(\widetilde{\boldsymbol{b}}^{\circ}|\boldsymbol{X},\boldsymbol{M})$, consists of three parts: the true coefficient $\boldsymbol{\beta}_1$, the individual-level bias $\widetilde{\boldsymbol{F}}\boldsymbol{\beta}_2$, and the aggregated-level bias $\widetilde{\boldsymbol{H}}^{\circ}\boldsymbol{\beta}_2$. These three parts are separable and additive.

Second, given $\boldsymbol{U}$ is essentially correlated with $\boldsymbol{X}_2$, the aggregated-level bias is not independent of the individual-level bias. However, $\boldsymbol{M}$ is not involved in $\widetilde{\boldsymbol{F}}$ but solely contributes to $\widetilde{\boldsymbol{H}}^{\circ}$, therefore, $\widetilde{\boldsymbol{b}}^{\circ}$ will be biased differently for $\boldsymbol{\beta}_1$ when the individual observations are aggregated differently, *ceteris paribus*. Fotheringham and Wong (1991) had vividly endorsed this property. This is the also theoretical reasoning on why multiscalar analysis could fail if there is an omission error in the regression.

Third, it is hard to judge the signs and the magnitudes of biases on both levels; $\widetilde{\boldsymbol{b}}^{\circ}$ can either over- or under-estimate $\boldsymbol{\beta}_1$, according to simulation results (omitted here). However, if $\widetilde{\boldsymbol{b}}$ is unbiased for $\boldsymbol{\beta}_1$, $\widetilde{\boldsymbol{b}}^{\circ}$ is unbiased for $\boldsymbol{\beta}_1$, too.

4. What can we do?

Multiscalar analysis is not a panacea. Because of the MAUP, the decisions about the scale and aggregation do impact our (imperfect) inferences about the world.

However, there is side B of the story. Multiscalar analysis does not alleviate the bias of the coefficient estimates, but it does provide a potential path to make inference about the missing independent variable if the researcher has the full control of schemes of spatial configurations. In addition, when the goal of conducting a regression analysis is not to identify causality but to make predictions, the MAUP helps to partially harvest the predictability concealed in the missing independent variables via their multicollinearity with the remaining independent variables. These are exciting research topics that can advance the frontier of the discipline.

References

Fotheringham, A. S., and Wong, D. W.-S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A: Economy and Space, 23*(7), 1025-1044.

Greene, W. H. (2012). *Econometric Analysis* (7th ed.). Essex, England: Pearson Education Limited.

Hennerdal, P., and Nielsen, M. M. (2017). A multiscalar approach for identifying clusters and segregation patterns that avoids the modifiable areal unit problem. *Annals of the American Association of Geographers, 107*(3), 555-574.

Larsen, J. L. (2000). *The modifiable areal unit problem: A problem or a source of spatial information?* (9962420 Ph.D.), The Ohio State University, Ann Arbor. Retrieved from http://search.proquest.com/docview/304635509 ProQuest Dissertations & Theses Global database.

Openshaw, S. (1983). *The Modifiable Areal Unit Problem* (Vol. 38). Norwick: Geo Books.

Openshaw, S., and Taylor, P. J. (1981). The modifiable areal unit problem. In Wrigley, N. and Bennett, R. J. (Eds.), *Quantitative Geography: A British View* (pp. 60-69). London: Routledge and Kegan Paul.