Mathematical & Descriptive Model

Taller de Tesis - Universidad de San Andrés

Juan Segundo Zapiola

November 2024

1. Spatial Weight Matrix & Correlation Test

In spatial econometrics, assuming independence between observations is highly restrictive, as the observed independent values y for a given observation i are often statistically similar to those in nearby locations. This phenomenon is known as spatial dependence (Ahumada et al., 2018). To estimate it, a starting point is to define an appropriate weights matrix, $\mathbf{W}$. Based on Tobler (1970) "everything is related to everything else, but near things are more related than distant things", this weights matrix represents the independence and interactions strength between the spatial observations.

Since the community's proportion of land dedicated to improved seed varieties are more likely to be influenced by their nearby neighbour communities rather than the remotes ones, the inverse distance between communities' locations is usually used as weights to measure the proximity between communities and neighboring communities (Elhorst, 2014; LeSage and Pace, 2009). To construct the distance-based weight matrix $\mathbf{W_d}$, following the studies of Zheng et al. (2021); Fang and Richards (2018); Li et al. (2021), it is measured as

$$w_{ij} = \begin{cases} 0 & \text{if } \; i = j \\ 1/d_{ij} & \text{if } \; d_{ij} \leq D \\ 0 & \text{if } \; d_{ij} > D \end{cases} \tag{1}$$

where d_{ij} is the arc-distance between the centroids of community i and the neighbouring community j. D is the critical distance cut-off, it would be assumed that there is no spatial relationship between communities that are beyond the distance D from each other, as indicated in equation (1). The critical distance cut-off D to be used would be 100 km. An extra weight matrix would be defined for comparison to emphasize on immediate neighbours. This weight matrix is going to be based on rook continuity, which only observations that are adjacent to the focal observation in rook fashion are considered neighbors (Fang and Richards, 2018). Both matrices would be row standardised.

To test whether there is spatial dependence on the proportion of land dedicated to improved seed varieties, the Moran's I test would be employed for every year in the panel:

$$\text{Moran's I} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{S^2 \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \tag{2}$$

Where $S^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, y_i indicates the observed value in the ith communities, w_{ij} is the spatial weight of the link between two communities i and j (as defined above), and n refers to the total number of communities. In addition to this, a GWR estimation would be implemented to determine if there is spatial heterogeneity on improved seed adoption in Malawi.

2. Modeling spatial panel data

According to the theoretical framework, the proportion of land dedicated to improved seed varieties y in a community i can be explained by it's characteristics and it's information acquisition. This can be specified as a non-spatial pooled regression model:

$$ln(y_{it}) = \beta_1 X_{it} + \beta_2 G_{it} + u_{it} \tag{3}$$

where y is the proportion of land dedicated to improved seed varieties in a community i in the year t. X_{it} represents a matrix of the community's characteristics, and G_{it} a matrix of community's information variables. β_1 and β_2 are vectors of parameters to be estimated, and u_{it} refers to the error term.

Based on the literary review, the proportion of land dedicated to improved seed varieties y in a community i, is also affected by their neighboring communities proportion of land dedicated to improved seed varieties y_j, its characteristics X_j and information acquisition G_j. A general nested pooled model [1] that incorporates all types of spatial interaction effects takes the form,

$$ln(y_{it}) = \rho W ln(y_{it}) + \beta_1 X_{it} + \beta_2 G_{it} + W X_{it}\theta_1 + W G_{it}\theta_2 + u_{it}; \quad u_{it} = \lambda W u_{it} + \epsilon_{it} \tag{4}$$

the endogenous spatial lag is specified as Wy (substantive spatial dependence), the exogenous spatial lag is specified as WX, and the residual spatial lag is specified as Wu (residual spatial dependence).

To capture spatial correlation, one of the standard approaches in most spatial analysis is to start with a general model containing, nested within it as special cases, a series of simpler models that ideally should represent all the alternative economic hypothesis requiring consideration (Elhorst, 2014). LeSage and Pace recommend using the Spatial Durbin Model (SDM) as a starting point, as it nests the largest number of simpler models. Additionally, Elhorst suggests that if the SDM cannot be reduced to a simpler model, one should also consider starting from the Spatial Durbin Error Model (SDEM) to compare the predictive performance of both models. Following the strategy from a general to a specific model (GETS), the Likelihood Ratio Tests (LR) would be used. This is because, when estimating more complex models, the subsequent models will be nested within them. LR tests allow for the comparison of the differences in log-likelihoods between the more complex model and the simpler one, helping to determine which is more appropriate. In addition, as (Elhorst, 2014) recommends, a fixed effect model would be applied as spatial units are located in unbroken study areas, in our case regions of Malawi. To estimate the spatial models Maximum Likelihood (ML) would be applied as estimates using Ordinary Least Squares (OLS) are biased and inconsistent.

[1] This Cliff-Ord model has an identification problem, making it impossible to estimate all the spatial coefficients. The most complete model can only incorporate two over the three possible spatial effects. It is shown as a reference model too illustrate all the spatial interaction effects.

Following the analysis of an static panel spatial model, a dynamic time-spatial panel model is constructed by adding the spatial-time lag value of the proportion of land dedicated to improved seed varieties:

$$\ln(y_{it}) = \gamma \ln(y_{i,t-1}) + \rho W \ln(y_{it}) + \eta W \ln(y_{i,t-1}) + \beta_1 X_{it} + \beta_2 G_{it} + W X_{it}\theta_1 + W G_{it}\theta_2 + u_{it} \tag{5}$$
$$u_{it} = \lambda W u_{it} + \epsilon_{it}$$

This model will assess how relevant time and spatial-time lag value of the proportion of land dedicated to improved seed varieties is. In presence of a dynamic model, the short and long run marginal effects can be obtained.

Referencias

Ahumada, H. A., Gabrielli, M. F., Herrera Gomez, M. H., and Sosa Escudero, W. (2018). Una nueva econometría: Automatización, big data, econometría espacial y estructural.

Elhorst, J. P. (2014). Linear spatial dependence models for cross-section data. *Spatial econometrics: From cross-sectional data to spatial panels*, pages 5–36.

Fang, D. and Richards, T. J. (2018). New maize variety adoption in Mozambique: a spatial approach. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie*, 66(3):469–488.

LeSage, J. and Pace, R. K. (2009). *Introduction to spatial econometrics*. Chapman and Hall/CRC.

Li, F., Feng, S., Lu, H., Qu, F., and D'Haese, M. (2021). How do non-farm employment and agricultural mechanization impact on large-scale farming? a spatial panel data analysis from Jiangsu Province, China. *Land Use Policy*, 107:105517.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, 46(sup1):234–240.

Zheng, H., Ma, W., and Li, G. (2021). Learning from neighboring farmers: Does spatial dependence affect adoption of drought-tolerant wheat varieties in China? *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie*, 69(4):519–537.