Índice

Índice de Figuras

Índice de Tablas

1 Introducción

2 Breve revisión de la literatura

3 Metodología

La autocorrelación espacial puede definirse como la similitud de valores en localizaciones próximas, ya sea positiva o negativa, su existencia implica que una muestra de datos contiene menor información que una muestra no correlacionada, formalmente:

$$Cov(y_i, y_j) = E(y_i, y_j) - E(y_i)(y_j) \neq 0 \quad \forall i \neq j$$

donde y_i, y_j son observaciones de una variable aleatoria en la localización i y j en el espacio, es decir, todo par (i, j) posee información geográfica específica media por latitud y longitud.

Sokal & Oden (1978) argumentaron que el análisis de autocorrelación espacial prueba si el valor observado de una variable nominal, ordinal o de intervalo en localidades independiente de los valores de esa misma variable en las localidades vecinas.

Entonces se puede llegar a la conclusión de que, si el valor de una o varias variables en una ubicación son similares a los valores de dichas variables en ubicaciones cercanas, entonces se dice que el patrón en conjunto exhibe una autocorrelación espacial positiva.

Por el contrario, se dice que existe autocorrelación espacial negativa cuando las observaciones que están cerca en el espacio tienen a ser más diferentes en los valores de las variables que las observaciones que están más separadas.

En los análisis de autocorrelación espacial se necesita una medida de contigüidad que podemos definirla de manera general como una relación de vecindad, estas pueden ser de tres tipos, caso de torre, caso de alfil y caso de la reina. Un aspecto crucial de la

definición de la autocorrelación espacial es la determinación de las ubicaciones cercanas, es decir, aquellas ubicaciones que rodean un punto de datos que podría considerarse que influyen en la observación en ese punto de datos.

Sin embargo, la determinación de este vecindario tiene un cierto grado de arbitrariedad. El número de observaciones en el vecindario establecido para cada ubicación puede expresarse mediante una matriz de ponderaciones W, que describe la conectividad entre n unidades que se encuentran localizadas en un espacio bidimensional.

$$\begin{bmatrix} W_{11} & W_{12} & \dots & W_{1n} \\ W_{21} & W_{22} & \dots & W_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ W_{n1} & W_{n2} & \dots & W_{nn} \end{bmatrix}$$

donde n representa el número de ubicaciones. La entrada en la fila i y columna j, denotado como W_{ij} corresponde al par (i, j) de ubicaciones. Los elementos diagonales de la matriz son cero, por convenio, mientras que los demás elementos donde $i \neq j$ toman valores distintos de cero cuando las ubicaciones se consideran vecinas.

Nosotros nos basaremos en la distancia euclídea (tipo reina), este método considera vecinas dos áreas si cumplen cierta propiedad referente a la distancia que las separa. Nos referimos a la distancia entre dos regiones como la distancia entre sus puntos representativos (centroides).

Al establecer un centroide en cada provincia se detectó que en el caso de contigüidad tipo reina la provincia de tierra del fuego quedaba desconectada del resto, esto es

porque esté criterio no obliga a los polígonos que se encuentran aislados geográficamente a relacionarse o conectarse con los demás. Una solución común a este problema es restringir la estructura contigua a los $k-$vecinos más cercanos, en nuestro caso $k = 4$, y por lo tanto excluir las 'islas' (áreas que por no estar a una distancia d de otra área se podría decir que no tiene vecinos) y forzar a cada unidad de área a tener el mismo número k de vecinos. Formalmente

$$W_{ij} = \begin{cases} 1 & \text{Si el centroide de } j \text{ es uno de los } k \text{ centroides más cercanos al de } i \\ 0 & \text{c.c} \end{cases}$$

Una vez definida la matriz de contactos, pueden utilizarse diferentes estadísticos univariantes que permiten detectar autocorrelación espacial. Los mismos pueden clasificarse como medidas de dependencia globales o locales. Las medidas globales utilizan la información completa del conjunto de datos con el fin de obtener un valor promedio para todo el espacio geográfico. Al resumir en un único valor toda la información, no es posible detectar la variabilidad de la dependencia ni la localización de estos patrones. Por su parte, las medidas locales examinan la autocorrelación espacial en un subconjunto de datos.

A la hora de estudiar los datos espaciales, está puede estar presente en variables explicativas, variable dependiente o en los residuos del modelo. Cuando la dependencia se encuentra en la variable dependiendo los modelos se denominan modelos de retardo espacial mientras que si está en los residuos se denominan modelos de error espacial. Por otro lado, cuando se presenta en las variables explicativas se llaman modelos de regresión cruzada o modelos X espacialmente retardados. Entonces, lo que primero se hace

es armar el modelo como si fuera una regresión lineal y ver dónde está la autocorrelación espacial y ver cuál es el modelo que mejor se ajusta.

Es por ello que presentaremos la prueba de Hausman (1978), que permite decidir entre un modelo en el cual los efectos individuales no se correlacionan con las variables explicativas, este test determina qué método de estimación utilizar.

La prueba de especificación de Hausman, puede aplicarse para probar el modelo de efectos aleatorios contra el modelo de efectos fijos. En nuestro caso, esta prueba se construye midiendo la brecha (ponderada por una matriz de varianza de covarianza) entre las estimaciones producidas por los estimadores dentro (modelo de efectos fijos) y GLS (modelo de efectos aleatorios) de los cuales se sabe que uno de los dos es convergente independientemente de la hipótesis formulada sobre la correlación entre variables y características inobservables, mientras que el otro (GLS) no converge en el único caso en el que esta hipótesis no se verifica. Por tanto, una diferencia significativa en ambas estimaciones implica una mala especificación del modelo de efectos aleatorios.

Mutl & Pfaffermayr (2011) han demostrado que estas propiedades siguen siendo válidas en un entorno espacial al reemplazar cada estimador dentro y GLS por su "análogo" espacial (teniendo en cuenta los términos de autocorrelación espacial). La prueba robusta de Hausman de autocorrelación espacial está escrita:

$$S_{hausman} = NT(\hat{\beta}_{MCG} - \hat{\beta}_{Within})' \left(\sum_{Within} - \sum_{MCG} \right)^{-1} (\hat{\beta}_{MCG} - \hat{\beta}_{Within})$$

donde $\hat{\beta}_{MCG}$ y $\hat{\beta}_{Within}$ son los parametros estimados por GLS y within respectivamente, $\sum_{Within} - \sum_{MCG}$ se corresponden a las matriz de varianza y covarianzas de los dos

estimadores.

A su vez, también presentaremos otros test que nos ayudaran a elegir la especificación más adecuada para estimar nuestro modelo, como los test I de Morán y Multiplicadores de lagrange. Por lo que partimos de un modelo estático (el más simple posible):

$$y = X\beta + \mu$$

$$\mu \sim (0, \sigma^2 I_n)$$

siendo la variable dependiente y un vector de dimensión $(n \times 1)$, X es una matriz de variables expicativas, incluyendo una constante, de orden $(n \times k)$, β es un vector de parámetros desconocidos de orden $(k \times 1)$ y μ es el término de error de dimendición $(n \times 1)$.

La presencia de estructura espacial en el modelo anterior puede contrastarse en base a estadísticos simples que utilizan resultados de estimación por mínimos cuadrados ordinarios (MCO). Uno de estos contrastes es el test I de Moran que se aplica a los residuos del modelo, sugerido por Cliff & Ord (1981):

$$I = \frac{n}{S_0} \frac{\hat{u}' W \hat{u}}{\hat{u}' \hat{u}}$$

donde $\hat{u}$ es el vector de residuos MCO, n es el número de observaciones y S_0 es la suma de todos los elementos de W.

La hipótesis nula del contraste es no autocorrelación espacial. El problema con esté test es que el rechazo de la hipotesis nula no brinda información sobre el posible modelo a especificar dado que la hipotesis alternativa es general y no da una guía sobre el tipo

de estructura espacial.

Como complemento, se emplea también el test de multiplicadores de Lagrange, LM, donde la hipótesis alternativa se encuentra bien definida y restringuida, estableciendo un modelo de error autoregresivo asumiendo que el término de error del modelo inicial se comporta de la siguiente forma:

$$u = \rho W \mu + \epsilon$$

donde ρ es el parámetro espacial autoregresivo, W es una matriz de pesos espaciales no estocástica de orden $(n \times n)$ y ϵ es un vector de innovaciones con media nula y varianza constante 0, $(\sigma^2 I)$. Entonces, el test establece las siguientes hipótesis:

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

además su formula de calculo es:

$$LM_{error} = \frac{1}{T_1} \left(\frac{\hat{u}' W \hat{u}}{\hat{\sigma}^2} \right)^2 \overset{as}{\sim} \chi^2_{(1)}$$

Donde T_1 es igual a $tr[(W' + W)W]$ y $\hat{u}$ son los residuos y $\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{n}$

Una segunda hipótesis alternativa, planteando un modelo de rezago espacial (SLM), que incorpora un rezago espacial de la variable dependiente, Wy, como explicativa:

$$y = \lambda W y + X\beta + \mu$$

donde λ es un parámetro espacial autoregresivo, μ es un vextor de errores de dimensión

$(nx1)$ y $\mu \sim (0, \sigma^2 I_n)$. Las hipotesis que propone el test son las siguientes:

$$H_0 : \lambda = 0,$$

$$H_1 : \lambda \neq 0$$

con la siguiente estructura:

$$LM_{LAG} = \frac{1}{n} \frac{\left(\frac{\widehat{u}' W \widehat{u}}{\widehat{\sigma}^2}\right)^2}{J_{\lambda\beta}} \overset{as}{\sim} \chi^2_{(1)}$$

con $J_{\lambda\beta} = \frac{1}{n\sigma^2}[(WX\beta)' M (WX\beta) + T_1 \sigma^2]$, con $M = I - (X'X)^{-1} X'$.

Según Herrera Gómez (2017) estos contrastes LM presentan como inconveniente que son sensibles a diferentes tipos de errores de especificación. Por ejemplo, el LM_{ERROR} detecta autocorrelación espacial debido a la presencia de un rezago espacial de la variable endógena (Wy), y lo mismo puede decirse del LM_{LAG}, que brinda falsos positivos cuando el término de error contiene un rezago espacial (Wu). Es por ello que Anselin et al. (1996) proponen dos nuevos multiplicadores de Lagrande que se caracterizan por ser más robustos a los errores de especificación.

EL LM^*_{ERROR} analiza la falta de correlación en los residuos, siendo robusto a la omisión del término Wy:

$$LM^*_{ERROR} = \frac{\left[\left(\frac{\widehat{u}' W \widehat{u}}{\widehat{\sigma}^2}\right) - \left(\frac{\widehat{u}' W \widehat{u}}{\widehat{\sigma}^2}\right)\right]^2}{n\widehat{J}_{\lambda\beta} - T_1} \overset{as}{\sim} \chi^2_{(1)}$$

Y el LM^*_{LAG} permite detectar la autocorrelación espacial en presencia de estructura espacial en el término de error:

$$LM^*_{LAG} = \frac{\left(\frac{\widehat{u}' W \widehat{u}}{\widehat{\sigma}^2}\right) - T_1 (\widehat{J}_{\lambda\beta})^{-1} \left(\frac{\widehat{u}' W y}{\widehat{\sigma}^2}\right)}{n\widehat{J}_{\lambda\beta} - T_1} \overset{as}{\sim} \chi^2_{(1)}$$

Estos contrastes permiten incorporar evariantes espaciales en base al rechazo o no de cada una de las hipótesis nulas. en base a la siguientes estrategía:

- Si I de Moran rechaza $H0$ hay evidencia a favor de inclusión de elementos espaciales.
- Si LM_{ERROR} y LM^*_{ERROR} rechazan $H0$ hay evidencia a favor de un modelo de error espacial (SEM).
- Si LM_{LAG} y LM^*_{LAG} rechazan $H0$ hay evidencia a favor de un modelo de rezagoespacial (SLM).
- Si no se rechaza $H0$ bajo ninguno de los contrastes, entonces hay evidencia a favor delmodelo lineal general no espacial.
- Si ambos contrastes robustos, LM_{ERROR} y LM^*_{LAG}, rechazan $H0$ entonces se deberán incorporar elementos espaciales en la parte sistemática (Wy) y aleatoria (Wu).

El modelo a implementar es el espacial autoregresivo, formalmente se presenta la siguiente manera:

$$y_{it} = \rho \sum_{i \neq j} w_{ij} y_{jt} + x_{it}\beta + \alpha_i + \mu_{it}$$

donde $u_{it} \overset{iid}{\sim} N(0, \sigma^2)$. La interacción espacial aquí se modela a través de la introducción de la variable dependiente espacialmente rezagada $\sum_{i \neq j} w_{ij} y_{jt}$. Al igual que en los modelos de sección transversal, la introducción de este variable conlleva efectos

secundarios globales: en promedio, el valor de y en el tiempo t para la observación i es explicado no solo por los valores de las variables explicativas de esta observación, sino también por aquellos asociado con todas las observaciones (vecino i o de otro tipo). También está en juego un efecto de derrame espacial global: un choque aleatorio en una observación i en el tiempo t afecta no solo el valor de y de esta observación en el mismo período, sino que también tiene un efecto en la valores de y de otras observaciones.

4 Datos

El estudio se realizó con datos de panel que comprenden un período desde 1993 al 2018 inclusive, correspondientes a las provincias argentinas y la Ciudad de Buenos Aires. Para poder analizar la influencia que tienen el tamaño de la población y el número de universidades en el producto bruto geográfico (PBG).

Contamos con un total de 624 observaciones de cada variable. En la tabla 1 se puede observar un resumen de las mismas. Podemos ver que el mínimo registrado para la población medida en miles de persones, es de 76.07 correspondiente a la provincia de Tierra del Fuego en el año 1993, por otra parte el máximo se corresponde la provincia de Buenos Aires para el año 2018 y el promedio de personas para el territorio es 1627.63.

Por otra parte, el PBG muestra una disparidad bastante notable dado que el presenta un minimo de 1.119 millones de pesos correspondiente a la provincia de Catamarca y un maximo de 2.964.595 para la Bueno Aires.

A su vez, para el número de universidades observamos que el mínimo es una universidad nacional, esto se da para las provincias de Corrientes, Catamarca, Jujuy, La Pampa, Misiones, Salta, San Juan y Tucumán. Y un máximo de nueve en la Ciudad de Buenos Aires.

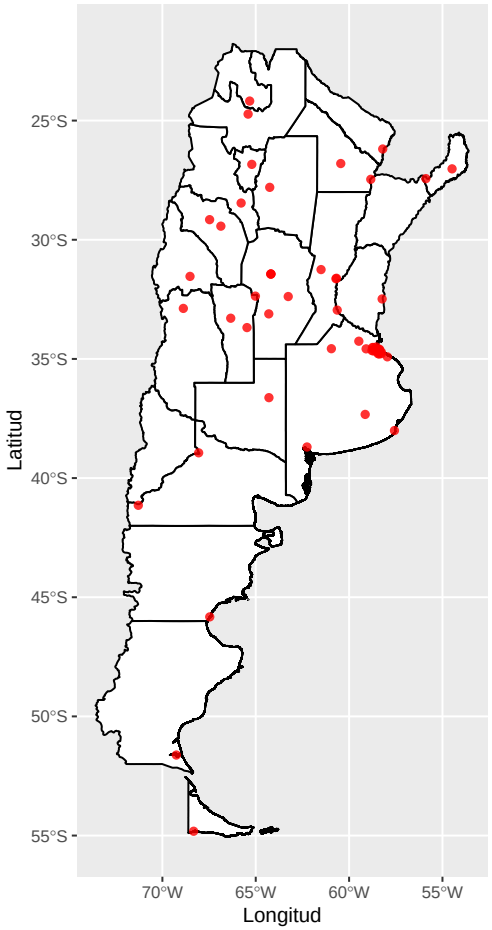
Tabla 1: Estadística Descriptiva

Variables	PBG	Población	#Universidades
Minimo	1.119	76.07	1
1s Quartil	3.830	424.26	1
Mediana	1.2610	748.77	2
Media	8.0130	1627.63	2.479
3rd Quartil	5.4836	1324.40	3
Máximo	2.964.595	17196.4	9

Fuente: Elaboración propia.

En la figura 1, se esboza la distribución espacial de las universidades nacionales en el territorio. Vemos que la mayor cantidad está ubicada en la provincia de Buenos Aires, esto podemos relacionarlo con la distribución de la población, ya que antes mencionamos que está provincia es la que más población concentra, seguida por la provincia de Córdoba.

Figura 1: Distribución geografica de las universidades.



Fuente: Elaboración propia.

5 Resultados

Nuestra aplicación como se mensionó anteriormente consiste en determinal la relación que tienen aumento en la población y el número de universidades de cada provincia,

con el producto bruto regional. Entonces, la especificación inicial está dada por:

$$PBG_{it} = b_0 + b_1\ POB_{it} + b_2\ \#universidades_{it} + \epsilon_{it}$$

donde b_0, b_1, b_2 son los parámetros desconocidos a estimar y ϵ_{it} es un término de error para el cual $i.i.d.$ inicialmente suponga que $\epsilon_{it} \sim N(0, \sigma^2)$.

Para tener en cuenta los efectos espaciales es necesario estimar la especificación aumentada por un término autorregresivo espacial (Fingleton, 2000), vinculando un error autorregresivo espacial modelo:

$$PBG_{it} = b_0 + b_1\ POB_{it} + b_2\ \#universidades_{it} + \epsilon_{it}$$

$$\epsilon_{it} = \alpha_i + \lambda \sum_{i \neq j} w_{ij}\epsilon_{jt} + v_{it}$$

Para seleccionar la especificación más adecuada, partimos del modelo sin autocorrección espacial, e implementamos la prueba de Hausman y las pruebas del multiplicador de Lagrange a los residuos del los modelos. En la tabla 2, muestra los resultados de la estimación donde en la columna (1) se observa un modelo de datos agrupados (pooled), mientras que las columnas (2) y (3) tienen en cuenta los datos no observados heterogeneidad individual, respectivamente, a través de efectos fijos y efectos aleatorios. (escribir alguna explicación breve sobre las estimaciones).

Por otro lado la prueba estándar de Hausman y su versión robusta para la detección de autocorrelación espacial de errores que se presenta en la tabla 3, conduce al rechazo de la hipótesis nula sobre la ausencia de correlación entre efectos y variables explicativas. Por lo que se elige un modelo de efectos fijo.

Tabla 2: Estimación sin considerar autocorrelación espacial

Model	pooled (1)	fixed effets(within) (2)	random effects (GLS) (3)
b_0 (intercep)	4.418634***	-	-3.422275**
log(Pob)	0.682238***	8.712956***	1.842667***
#Universidades	0.264833***	0.106046***	0.273810***
Observaciones	624	624	624
R^2 ajustado	0.47102	0.72179	0.332

***

p<0.01, ** p<0.05, * p<0.1. Fuente: Elaboración propia.

Tabla 3: Prueba de Hausman

Modelo	Pooled(1)	fixed effects(within) (2)	random effects (GLS) (3)
Estadistico chi 2	1316.3***	35.043***	17.048***

***

p<0.01, ** p<0.05, * p<0.1. Fuente: Elaboración propia.

Por otro lado, en la tabla 4, se exponen los resultados de las pruebas de los multiplicador de Lagrange (LM) en un modelo de efectos fijos, los cuales tienden a favorecer una especificación SAR del modelo.

Los estadisticos de la prueba para tomar la autocorrelación espacial por SAR (Prueba 1) o SEM (Prueba 2) confirman el rechazo de la hipótesis de que estos dos términos (tomados independientemente) son nulos, la lectura simultánea nos permite concluir sobre la especificación más adecuada para tener en cuenta la autocorrelación espacial. Para concluir de una manera más creíble, se utilizan pruebas robustas en la presencia

de la especificación alternativa de autocorrelación espacial (Pruebas 3 y 4).

La versión robusta de LM_{ERROR} no es significativa (Prueba 4) mientras que LM_{LAG} si lo es (Prueba 3). Por lo tanto, es conveniente estimar un modelo Fixed-effect SAR. (poner en pie de página tal vez: En algunos casos, estas dos últimaspruebas robustas no permiten discriminar entre un SAR y un SEM. Son posibles varias posibilidades. El primero consiste en estimar un modelo que contiene ambos términosespaciales (SARAR).))

Tabla 4: Test LM y LM*

LM_{LAG} (prueba 1)	LM_{Error} (prueba 2)	LM^*_{LAG} (prueba 3)	LM^*_{Error} (prueba 4)
807.74***	541.03***	267.46***	0.74516

*** p<0.01, ** p<0.05, * p<0.1. Fuente: Elaboración propia.

Entonces, teniendo en cuenta las concluciones de los test que ayudan a mejorar la especificación del modelo tenemos los resultados del mismo en la tabla 5. La columna (1) muestra el modelo de datos agrupados, mientras que en la columna (2) se exponen los resultados del modelo de efecto fijos con diferencias en la espacificación del término de error. Podemos observar que el coeficiente de autocorrelación es positivo y significativo.

En cuanto a los coeficientes estimados, llegamos a la conclución de que tanto log(POB) y *#universidades* son significativos, por lo que el aumento de la población tiene un impacto positivo en el pbg, al igual que la creación de nuevas universidades, pero en menor medida.

Tabla 5: Modelo

Model	pooled (1)	fixed effects(MV) (2)
b_0 (intercep)	6.408201***	-
log(Pob)	0.863431***	0.0718612*
#universidades	0.068693***	0.0160614***
λ	-	0.9770277***
Observaciones	624	624

*** p<0.01, ** p<0.05, * p<0.1. Fuente: Elaboración propia.

6 Conclusiones preliminares

Bibliografía

Cliff, A., & Ord, J. (1981). Spatial processes: Models and applications. Pion: London, UK.

Fingleton, B. (2000). Spatial econometrics, economic geography, dynamics and equilibrium: A "third way"? Environment and Planning A, 32(8), 1481–1498.

Hausman, J. A. (1978). Specification tests in econometrics. Econometrica: Journal of the Econometric Society, 1251–1271.

Herrera Gómez, M. (2017). Fundamentals of applied spatial econometrics. CONICET-IELDE, National University of Salta.

Mutl, J., & Pfaffermayr, M. (2011). The hausman test in a cliff and ord panel model. The Econometrics Journal, 14(1), 48–76.

Sokal, R. R., & Oden, N. L. (1978). Spatial autocorrelation in biology. Biological Journal of the Linnean Society, 10(2), 199–228.