Ingrid Acevedo Bohórquez
Ermilson Velásquez Ceballos

Resumen

El presente artículo es parte del trabajo realizado en el proyecto de Investigación "Análisis Exploratorio de Datos Espaciales y el Índice de Moran", el cual es financiado con fondos de la Universidad EAFIT para el año 2008. En este trabajo se presenta una descripción de los efectos espaciales que se pueden encontrar cuando se trabaja con datos georreferenciados, además se hace un desarrollo formal introductorio relacionado con el estadístico I de Moran, uno de los test más utilizados para contrastar la autocorrelación espacial. Se presenta también un caso de aplicación en el cual se muestra la importancia del análisis exploratorio de datos espaciales (AEDE), en el proceso de análisis de fenómenos que estén enmarcados en un contexto regional.

Palabras Clave: Econometría espacial, Analisis Exploratorio de Datos Espaciales.

Abstract

This work presents we present a description of the spatial effects that can found when working with referred geographically data, in addition it presents a formal development of the Moran´s I statistic, which is one the most used test for proving existence of spatial autocorrelation. We include a case of application where we show the importance of the Exploratory Spatial Data Analysis – ESDA in the process of studying phenomena that are framed in a regional context.

Keywords: Spatial Econometrics, Exploratory Spatial Data Analysis

JEL Classification: R12, C40.

Algunos conceptos de la econometría espacial y el análisis exploratorio de datos espaciales

Ingrid Acevedo Bohórquez[*]
Ermilson Velásquez Ceballos[**]

1. Introducción

El análisis exploratorio de datos espaciales - AEDE, constituye una disciplina reciente que ha adquirido una especial importancia debido principalmente al avance de la tecnología en las comunicaciones y la globalización de la economía.

Los sucesos que ocurren en una ubicación específica tienen repercusiones sobre sus vecinos directos e incluso sobre otros, aparentemente remotos. En el estudio de cualquier fenómeno de carácter social o económico la ubicación geográfica de los agentes constituye un aspecto importante dentro de la especificación de los modelos econométricos, ya que puede existir algún efecto espacial, que de no ser incorporado en la especificación, podría afectar la validez del modelo. Ante esta realidad y gracias al desarrollo tecnológico de los sistemas de georreferenciación de datos, surge la necesidad de contar con herramientas apropiadas para el procesamiento, descripción y análisis de la información ya que los métodos tradicionales de la estadística descriptiva no tienen en cuenta la localización geográfica de los datos.

Teniendo en cuenta que la econometría tradicional no ha incorporado el efecto de dichas circunstancias y que la estadística espacial se

Fecha de recepción: 1 de agosto de 2008. Fecha de aceptación: 22 de septiembre de 2008.

[*] Ingeniera Civil, Magister en Matemáticas Aplicadas de la Universidad EAFIT, iacevedo@eafit.edu.co

[**] Doctor en Ciencias Matemáticas, profesor de tiempo completo de la Universidad EAFIT, evelas@eafit.edu.co

ocupa de otro tipo de problemas, ha surgido una disciplina a la cual se le ha dado el nombre de econometría espacial. Según Luc Anselin, uno de sus principales investigadores, "las actividades como la estimación de modelos espaciales de interacción, el análisis estadístico de la función de densidad urbana y la implementación empírica de modelos econométricos regionales, podrían ser considerados econometría espacial" (Anselin, 1988)

Al utilizar información georreferenciada surge el interrogante sobre si se encuentra presente algún tipo de dependencia espacial entre los datos. Esta dependencia se denomina autocorrelación espacial y es el más importante de los efectos espaciales. Para contrastar su presencia el estadístico más utilizado fue propuesto por Moran en los años cincuenta y a partir de el se han diseñado nuevas propuestas.

El objetivo principal del presente artículo es presentar una introducción al desarrollo formal relacionado con el índice I de Morán y realizar una aplicación utilizando información sobre la variable número de homicidios en los municipios del departamento de Antioquia.

Antes de comenzar describiendo cualquier método estadístico de análisis de datos espaciales es necesario definir que se debe entender como datos espaciales. Un dato espacial puede ser definido como la observación de una variable asociada a una localización del espacio geográfico.

Cuando se tienen observaciones georreferenciadas, se deben utilizar herramientas que permitan detectar ciertas características dentro de los datos, como son tendencia, valores atípicos, esquemas de asociación y dependencia espacial, concentración espacial o puntos calientes/fríos, entre otros.

Aunque en la actualidad se tiene gran cantidad de información georreferenciada, estos datos suelen ser tratados con herramientas del análisis de series temporales (o de corte transversal, no espacial), sin usar técnicas adecuadas para el análisis estadístico espacial.

Los métodos que permiten extraer dichas características de los datos georreferenciados se conocen con el nombre de análisis explo-

ratorio de datos espaciales (AEDE) y se conciben como una disciplina dentro del análisis estadístico más general, diseñada para el tratamiento específico de los datos geográficos. El AEDE se utiliza para identificar relaciones sistemáticas entre variables, o dentro de una misma variable, cuando no existe un conocimiento claro sobre su distribución en el espacio geográfico (Chasco Yrigoyen, 2006)

Al tratar de analizar los datos georreferenciados surge la pregunta. ¿Por qué varían las relaciones sobre el espacio? La razón más simple es que existen variaciones espaciales en las relaciones observadas debidas a variaciones muestrales aleatorias. Una segunda razón puede atribuirse a que las relaciones en si pueden ser diferentes a través del espacio, tal vez porque existen variaciones espaciales en las actitudes o preferencias de la población o existen asuntos administrativos, políticos o de otros contextos que producen respuestas diferentes ante el mismo estimulo. Una tercera razón puede ser la omisión o una representación funcional incorrecta de una o más variables relevantes para la explicación del modelo (Haining, 2003).

El objetivo principal del análisis exploratorio de datos espaciales esta relacionado con la identificación de excepciones locales o tendencias generales, ya sea en los datos o en las relaciones.

El Análisis Exploratorio de Datos Espaciales debe ser la etapa inicial de cualquier estudio econométrico que involucre datos georreferenciados.

2. Efectos Espaciales

En el presente numeral se hace una breve descripción de los efectos espaciales que pueden presentarse cuando se trabaja con datos georreferenciados, estos efectos son los que impiden que los métodos de la econometría estándar sean una buena herramienta para su modelación.

La información obtenida para uso de la ciencia regional posee características que provienen de su ubicación geográfica. Esta información posee características que constituyen los denominados efectos espaciales, los cuales pueden ser divididos en dos tipos: dependencia espacial y heterogeneidad espacial.

De los dos, la dependencia espacial o la autocorrelación espacial es la más conocida. La dependencia espacial significa la ausencia de independencia que con frecuencia está presente entre observaciones en conjuntos de datos. Esta dependencia puede ser expresada según la primera ley de la geografía de Tobler (1970), en la cual "todo esta relacionado con todo lo demás, pero las cosas cercanas están más relacionadas que las cosas distantes".

El segundo tipo de efecto espacial, la heterogeneidad espacial, está relacionado con la ausencia de estabilidad en el comportamiento o las relaciones bajo estudio. Más precisamente, esto implica que los parámetros y formas funcionales varían con la ubicación y no son homogéneos en los conjuntos de datos.

2.1 Dependencia Espacial

La dependencia espacial en una colección de observaciones de datos muestrales se refiere al hecho de que una observación asociada con una localización la cual se puede denominar *i*, depende de otras observaciones asociadas con localizaciones $j \neq i$. Esto puede ser formalmente expresado como

$$y_i = f(y_j) \qquad i = 1,\ldots,n \qquad j \neq i$$

Donde cada observación de una variable y en $i \in S$ (con S como el conjunto que contiene todas las unidades espaciales de observación), está relacionada formalmente a través de la función *f*, con las magnitudes de la variable en otra unidad espacial en el sistema.

Esta simple expresión, por si misma, no es muy útil en una situación empírica, dado que resultaría un sistema inidentificable, con muchos más parámetros (potencialmente $N^2 - N$) que observaciones (N). Imponiendo una estructura en las relaciones funcionales involucradas en *f,* es decir, una forma particular para el proceso espacial, un número limitado de características de la dependencia espacial podría ser estimado y probado empíricamente.

Surge así el siguiente cuestionamiento. ¿Por qué se esperaría que los datos observados en un punto del espacio sean dependientes de

valores observados en otros puntos? Tres tipos de condiciones podrían conducir a una situación como la descrita.

La primera puede ser producto de los errores de medición para las observaciones en unidades espaciales contiguas. Ejemplos de estos problemas son: La delimitación arbitraria de las unidades espaciales de observación, las cuales no recogen adecuadamente el proceso generador de los datos muestrales.

La segunda está relacionada con problemas de agregación espacial, ya que en muchas situaciones encontradas en la práctica, la información es recolectada solo a una escala agregada, por lo tanto, podría haber muy poca correspondencia entre el alcance del fenómeno agregado bajo estudio y la delimitación de las unidades espaciales de observación.

Una tercera y tal vez más importante razón que podría causar dependencia espacial, proviene de la importancia del espacio como elemento estructural fundamental en explicaciones sobre el comportamiento humano y las actividades económicas. La ciencia regional esta basada en la premisa de que la ubicación y la distancia son fuerzas importantes en trabajos de geografía humana y actividades de mercado. Todas estas nociones han sido formalizadas en la teoría de la ciencia regional que se apoya en las nociones de interacción espacial, procesos de difusión y jerarquías espaciales.

2.2 Autocorrelación Espacial

Definida de manera simple, la autocorrelación espacial es la concentración o dispersión de los valores de una variable en un mapa. Dicho de otra manera, la autocorrelación espacial refleja el grado en que objetos o actividades en una unidad geográfica son similares a otros objetos o actividades en unidades geográficas próximas (Goodchild, 1987).

La dependencia espacial se produce cuando "el valor de la variable dependiente en una unidad espacial es parcialmente función del valor de la misma variable en unidades vecinas" (Flint, Harrower y Edsall, 2000). En el análisis de datos agregados geográficamente es frecuente

encontrar que los valores de las variables estén autocorrelacionados espacialmente o sean espacialmente dependientes. La diferencia entre autocorrelación espacial y dependencia espacial está, fundamentalmente, en el uso de las palabras y estriba en que el primer caso se refiere simultáneamente a un fenómeno y técnica estadística, y el segundo, a una explicación teórica (Vilalta y Perdomo, 2005).

2.3 Heterogeneidad Espacial

El término heterogeneidad espacial se refiere a la variación en las relaciones sobre el espacio. En el más general de los casos se podría considerar que una relación diferente se presente para cada punto en el espacio. Formalmente una relación lineal se representaría así:

$$y_i = X_i\beta_i + \varepsilon_i \qquad i = 1,..., n$$

La estimación de este tipo de modelos lleva implícito problemas relacionados con los grados de libertad, simplemente no se tienen suficientes datos con los cuales producir estimadores para cada punto en el espacio.

Para proceder con el análisis, se debe proveer una especificación para la variación sobre el espacio. Por lo tanto para llevar a cabo estimaciones e inferencias con soporte formal y asegurar la identificabilidad del modelo, es necesario imponer algunas restricciones a la expresión general.

En la literatura de ciencia regional y geografía económica, hay una amplia evidencia de la falta de uniformidad de los efectos del espacio. Muchos factores, tales como jerarquías del sitio central, la existencia de regiones lideres y regiones rezagadas, efectos de cosecha en el crecimiento urbano, etc.; son argumentos que requieren de estrategias que tomen en cuenta las características particulares de cada ubicación o unidad espacial. En el trabajo econométrico esto se puede llevar a cabo mediante la consideración de parámetros cambiantes, coeficientes aleatorios, o varias formas de cambio estructural.

Además, de esta falta de estabilidad estructural de varios de los fenómenos en el espacio, las unidades espaciales en observación son distantes de ser homogéneas. Por ejemplo, las unidades de los censos tienen diferente área y forma, los complejos urbanos tienen poblaciones

o niveles de ingreso desiguales, y las regiones tienen grados de desarrollo tecnológico diferentes. En la medida en que estos aspectos de heterogeneidad se reflejen en los errores de medición, podrían resultar en heteroscedasticidad.

A diferencia del caso de la dependencia espacial, los problemas causados por la heterogeneidad pueden ser, en la mayor parte, resueltos por los medios de las técnicas de la econometría estándar. Específicamente, métodos que pertenecen a parámetros cambiantes, coeficientes aleatorios e inestabilidad estructural pueden ser fácilmente adaptados para tomar en cuenta las variaciones en el espacio. Sin embargo, en muchos casos, la estructura espacial inherente en los datos puede llevar a procedimientos más eficientes. También, la compleja interacción, que resulta de la estructura espacial y de los flujos espaciales, podría generar dependencia en combinación con la heterogeneidad. En tal situación, el problema de distinguir entre dependencia espacial y heterogeneidad espacial es altamente complejo.

2.4 Heteroscedasticidad Espacial

La heteroscedasticidad consiste en la ausencia de estabilidad en la dispersión de un fenómeno, como sucede muchas veces con los residuos de una regresión y puede representarse como:

$$Var(u_i) = \sigma_i^2$$

donde σ_i^2 indica que la varianza de la perturbacion aleatoria es diferente para cada observacion muestral *i*.

3. Matriz de Contigüidad

Se denomina matriz de contigüidad o de conectividad al arreglo *W* donde tanto cada una de las filas como de las columnas representa una región en el espacio objeto de estudio. Esta matriz representa la relación que tiene cada una de las regiones con las demás regiones del espacio en estudio, tal como se vería en un mapa. Existen una infinidad de formas en que la matriz de contigüidad puede ser construida, la más sencilla es utilizando notación binaria, donde 1 representa la presencia de contigüidad espacial entre dos unidades y 0 la ausencia de contigüidad

espacial entre dos unidades. Una matriz construida de esta manera es simétrica.

Existe un desconcertante gran número de formas para definir la presencia o ausencia de contigüidad, algunas de las cuales se definen a continuación.

- Contigüidad de torre: en la cual se define $w_{ij} = 1$ para unidades que comparten un lado común con la región de interés a la izquierda, a la derecha, arriba o abajo.

	b	
b	**a**	**b**
	b	

- Contigüidad de alfil: en la cual se define $w_{ij} = 1$ para unidades que comparten un vértice común con la región de interés.

b		**b**
	a	
b		**b**

- Contigüidad de reina: Para unidades que comparten un lado en común o un vértice con la región de interés se define $w_{ij} = 1$.

b	**b**	**b**
b	**a**	**b**
b	**b**	**b**

Existen muchas más formas en las cuales se podría procesar la matriz de Contigüidad.

Una de las definiciones más usadas es la de la torre, con la cual se define la matriz de contigüidad de primer orden, ya que algunas veces solo es necesario establecer la localización de las unidades en el mapa que tienen bordes comunes con longitudes positivas.

En la construcción de matrices de contigüidad que se ajusten al problema de estudio puede que la matriz de contigüidad binaria no provea una buena especificación, por lo cual es necesario tener en mente otras posibles matrices. Algunas metodologías para la construcción de estas matrices de contigüidad se basan en la similaridad o disimilaridad existente entre los datos y las características inherentes a las variables en estudio. Para identificar esta similaridad se pueden utilizar las medidas de similaridad que reúnen propiedades de métrica, cuando se trata de datos cuantitativos, y los coeficientes de asociación, que son aquellos empleados para datos expresados en una escala nominal.

Existen además otras propuestas en las cuales se involucran variables socioeconómicas como el caso de Case, Rosen y Hines (1993) quienes proponen la utilización de una distancia económica, sugiriendo, por ejemplo, la definición de w_{ij} como $w_{ij} = \frac{1}{|x_i - x_j|}$, donde las x son observaciones de características socioeconómicas, como por ejemplo el ingreso per cápita; otra propuesta es la de Vaya, Lopez-Bazo y Moreno (1998), quienes proponen recoger en la matriz de contigüidad el grado de intercambio comercial entre las regiones analizadas.

La guía principal para seleccionar una definición sobre las demás, sería la naturaleza del problema que se pretende modelar y tal vez la información adicional disponible diferente de las observaciones muestrales. Ejemplo de esto puede ser la conexión entre regiones por medio de autopistas, conocimiento de relaciones comerciales entre regiones, etc., que nos ayudan a decidir cual de las definiciones de contigüidad se ajusta más a las regiones objeto de estudio.

En el trabajo aplicado a menudo se lleva a cabo la transformación de la matriz de contigüidad de tal forma que los elementos de cada uno de los renglones sumen uno. Esta nueva matriz se denomina la matriz de contigüidad de primer orden estandarizada $C = W_z$.

La motivación para la estandarización puede ser vista considerando lo que ocurre si multiplicamos la matriz estandarizada por un vector de observaciones en las diferentes regiones de una variable asociada y. Esta nueva matriz producto $y^* = Cy$ representa una nueva variable igual a la media de las observaciones de las regiones contiguas.

4. Análisis tradicional vs. Análisis espacial

El análisis exploratorio de datos tradicional se ha definido como una colección de técnicas para resumir las propiedades de los datos (estadística descriptiva), detectar patrones en los datos, identificar características inusuales en los datos y formular hipótesis desde los datos. Estas técnicas también se usan para examinar los resultados del modelo y proporcionar evidencia de si los supuestos del modelo son satisfechos. El conjunto de técnicas aplicadas son gráficas (gráficos y figuras) y/o numéricas (resúmenes cuantitativos de los datos).

De la misma manera, el análisis exploratorio de datos espaciales está constituido por diversas técnicas que permiten explorar los datos espaciales, resumir las propiedades espaciales de los datos, detectar patrones en los datos, formular hipótesis que se refieren a la presencia de fenómenos espaciales dentro de los datos, identificar casos o subconjuntos de casos que son inusuales dada su localización en el mapa, etc. A diferencia del análisis tradicional, en el análisis exploratorio de datos espaciales los mapas cobran especial importancia, ya que permiten responder preguntas tales como: ¿Donde se encuentran en el mapa los casos atípicos observados en el histograma? ¿Donde aparecen los valores de cierto atributo de una parte del mapa en el *scatterplot*? ¿Cuáles son los patrones espaciales y las asociaciones espaciales en este conjunto de datos?, etc.

5. Visualización de los datos

La visualización de datos espaciales emplea herramientas cartográficas (diferentes formas de presentación de mapas) y enlaces entre la cartografía y los diferentes gráficos estadísticos. Algunas utilidades de estos enlaces son entre otras:

- Enlace gráfico-mapa: permite identificar la localización en el área de estudio de casos particulares como los *outliers*.
- Enlace mapa-gráfico: permite identificar si un área determinada es distinta en términos de sus atributos.

- Enlace dinámico gráfico-mapa: permite moverse sobre un polígono del gráfico y obtener las regiones resaltadas en la ventana del mapa.
- Enlace dinámico mapa-gráfico: permite moverse sobre el mapa y obtener los casos resaltados en la ventana del gráfico, o calcular los estadísticos para los casos que caen dentro de las regiones resaltadas.
- El cartograma: es una presentación del mapa donde las áreas de las regiones son proporcionales a la población de cada región y puede proveer un mejor marco espacial del atributo analizado. (Haining, 2003).

Técnicas de Análisis Exploratorio de Datos Espaciales

Perspectiva Econometría Espacial	
Visualización de distribuciones espaciales	• *Box map,* • Histograma, • Análisis de la varianza exploratorio espacial
Visualización de asociación espacial global	• Gráficos del retardo espacial, • Mapa y *Scatterplot* de Moran
Visualización de asociación espacial local	• Mapas LISA, • *Outliers* en el *Scatterplot* de Moran
Asociación espacial multivariante	• *Scatterplot* multivariante de Moran
Heterogeneidad Espacial	• Mapa, • Histograma de Frecuencias, • Diagrama de Dispersión

Fuente: (Anselin, 1988)

La distribución no aleatoria de los fenómenos en el espacio tiene varias consecuencias para el análisis estadístico convencional. Por ejemplo, los parámetros estimados con base en muestras que no se

distribuyen aleatoriamente en el espacio son sesgados hacia valores prevalentes en las regiones favorecidas en el muestreo. Como resultado, muchos de los supuestos que se deben hacer sobre los datos antes de aplicar pruebas estadísticas no son válidos en estos casos.

Otro punto de vista al respecto es que la autocorrelación espacial introduce "redundancia" entre los datos, tal que cada unidad adicional aporta menos información nueva. Esto afecta el cálculo de los intervalos de confianza, etc. Tales efectos implican que sea recomendable evaluar el grado de autocorrelación espacial en un conjunto de datos espaciales antes de realizar cualquier análisis estadístico convencional basado en esta información.

6. El índice I de Moran

Entre las medidas de diagnóstico de autocorrelación disponibles están los estadísticos de conteo conjunto, el índice I de Moran, el índice C de Geary y la nube de variograma. Estas son técnicas que ayudan a detectar si entre las unidades espaciales hay o no autocorrelación espacial.

Indiscutiblemente, la herramienta más utilizada con este fin es el índice I de Moran, el cual es una adaptación de una medida de correlación no-espacial a un contexto espacial y se aplica normalmente a unidades espaciales donde hay disponibilidad de información en forma de razones o intervalos. Una de las especificaciones más utilizadas es

$$I = \frac{n}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i-\bar{y})(y_j-\bar{y})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}$$

Se puede analizar cada uno de los factores que intervienen en el índice.

Primero el numerador de la segunda fracción, es decir, $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i-\bar{y})(y_j-\bar{y})$, que se reconoce como el término de varianza, de

hecho es una covarianza. Los subíndices *i* y *j* se refieren a diferentes unidades o zonas espaciales en el estudio y y_i es el valor de la observación para cada una de ellas. Al calcular el producto de la diferencia de las observaciones de dos zonas con la media general $\bar{y}$ se determina hasta dónde varían las observaciones conjuntamente. Si tanto y_i como y_j están al mismo lado de la media, este producto es positivo; si por el contrario, están ubicados a lados diferentes de la media, el producto es negativo y el tamaño absoluto del valor resultante depende de qué tan cercanos sean los valores observados a la media general. Los términos de covarianza son multiplicados por w_{ij}, este es un elemento de la matriz de ponderaciones *W*. En el caso más sencillo,

$$w_{ij} = \begin{cases} 1 & \text{Si la región i y la región j son adyacentes} \quad \forall i \neq j \\ 0 & \text{En caso contrario} \end{cases}$$

Los demás elementos de la fórmula normalizan el valor de *I* respecto al número de zonas en consideración, el número de adyacencias del problema y el rango de valores en *y*.

El divisor $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$ introduce el número de relaciones en el mapa.

El factor $\frac{n}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{1}{V(y)}$ es en realidad una división por la varianza general del conjunto de datos. La ecuación del *I* de Moran puede escribirse en términos matriciales de forma más compacta como

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{y'Wy}{y'y}$$

Donde:

y: es el vector columna cuyas entradas son cada diferencia $(y_i - \bar{y})$.

El índice *I* de Moran y el índice Durbin y Watson son estructuralmente equivalentes al ser ambos un cociente de formas cuadráticas de

residuos de regresión, su diferencia está en las matrices de pesos que especifican los enlaces entre las observaciones.

La eficiencia y propiedades de los estimadores así como de otros estadísticos dependen en general de si los términos de error de un modelo son correlacionados espacialmente. De allí que sea importante probar la existencia de correlación espacial.

Moran en su artículo "Notes on Continuous Stochastic Phenomena", propuso un estadístico de prueba para el contraste de correlación entre unidades espaciales "adyacentes", cuyo desarrollo formal presentamos a continuación.

Sea X una matriz de variables *(m x n)* aleatorias x_{ij} $i = 1,..., m$ $j = 1,...,n$ independientes e idénticamente distribuidas

$$\bar{x} = \frac{1}{mn}\sum_{i}^{m}\sum_{j}^{n} x_{ij}$$

$$mn\bar{x} = \sum_{ij} x_{ij}$$

Sea $z_{ij} = x_{ij} - \bar{x}$

La covarianza entre las variables aleatorias z_{ij}, $z_{i,j+1}$ está dada por:

$$\begin{aligned} cov(z_{ij}, z_{i,j+1}) &= E\left[z_{ij} - E\left(z_{ij}\right)\right]E\left[z_{i,j+1} - E\left(z_{i,j+1}\right)\right] \\ &= E\left[z_{ij} z_{i,j+1}\right] \end{aligned}$$

Y por el principio de analogía

$$\hat{cov}(z_{ij}, z_{i,j+1}) = \frac{\sum_{i}^{m}\sum_{j}^{n-1} z_{ij} z_{i,j+1}}{m(n-1)}$$

Siguiendo un razonamiento similar podemos construir el coeficiente de autocorrelación espacial de primer orden de una manera natural, como se presenta en la siguiente expresión

$$r_{11} = \left(\frac{mn}{2mn-m-n}\right)\frac{\sum_{i}^{m}\sum_{j}^{n-1} z_{ij}z_{i,j+1} + \sum_{i}^{m-1}\sum_{j}^{n} z_{ij}z_{i+1,j}}{\sum_{ij} z_{ij}^2}$$

$$r_{11} = \left(\frac{mn}{2mn-m-n}\right)\frac{\sum_{i}^{m}\left(z_{i1}z_{i2} + z_{i2}z_{i3} + \cdots + z_{i,n-1}z_{in}\right) + \sum_{i}^{m-1}\left(z_{i1}z_{i+1,1} + z_{i2}z_{i+1,2} + \cdots + z_{in}z_{i+1,n}\right)}{\sum_{i}\left(z_{i1}^2 + z_{i2}^2 + \cdots + z_{in}^2\right)}$$

$$r_{11} = \left(\frac{mn}{2mn-m-n}\right)[(z_{11}z_{12} + z_{12}z_{13} + \cdots + z_{i,n-1}z_{in} + z_{21}z_{22} + z_{22}z_{23} + \cdots + z_{2,n-1}z_{2n} + \cdots + z_{m1}z_{m2} + z_{m2}z_{m3} + \cdots$$
$$+ z_{m,n-1}z_{mn}) + (z_{11}z_{21} + z_{12}z_{22} + \cdots + z_{1n}z_{2n} + z_{21}z_{31} + z_{22}z_{32} + \cdots + z_{2n}z_{3n} + \cdots + z_{m-1,1}z_{m1}$$
$$+ z_{m-1,2}z_{m2} + \cdots + z_{r-1,n}z_{mn})]/(z_{11}^2 + z_{12}^2 + \cdots + z_{in}^2 + z_{21}^2 + z_{22}^2 + \cdots + z_{2n}^2 + z_{m1}^2 + z_{m2}^2 + \cdots + z_{mn}^2)$$
$$r_{11} = \left(\frac{mn}{2mn-m-n}\right)I$$

Definición: El índice *I* de Moran se define a partir de la construcción anterior como

$$I = \frac{\sum_{i}^{m}\sum_{j}^{n-1} z_{ij}z_{i,j+1} + \sum_{i}^{m-1}\sum_{j}^{n} z_{ij}z_{i+1,j}}{\sum_{ij} z_{ij}^2}$$

7. Caso de Aplicación

En los estudios de fenómenos sociales la ubicación geográfica juega un papel relevante que debe ser tenida en cuenta al construir modelos en los cuales intervengan variables relacionadas con la ubicación geográfica. Estudios de este tipo de fenómenos que tengan en cuenta de manera explícita el concepto de territorio son muy escasos en Colombia. Por lo tanto se considera que realizar aplicaciones con datos de Colombia es un paso muy importante para futuros trabajos.

A continuación se presenta un caso de aplicación del Análisis Exploratorio de Datos espaciales con los datos suministrados por la Secretaria de Gobierno del Departamento de Antioquia.

Los resultados obtenidos fueron presentados en el informe de la investigación "Análisis de los Indicadores de Vida en el departamento de Antioquia 2001-2007" , estudio realizado por el Centro de Análisis Político de la Universidad EAFIT, bajo la dirección del Dr Jorge Giraldo.

La variable a analizar es el número de homicidios en cada uno de los 125 municipios que conforman el departamento. El análisis se realizó para cada uno de los años 2001 a 2006, pero en el presente artículo solo se presentan los resultados del año 2001.

Al realizar esta propuesta utilizando de manera explícita la información geográfica, se muestra la importancia que tiene el concepto de región para análisis con los datos de Colombia, esto se ve reflejado en el hecho de que a través del estudio se refleja que los fenómenos de violencia en Colombia no son procesos que se produzcan de manera aleatoria sino que obedecen a patrones.

7.1 Definición de variables

X_i : Es el número de homicidios en el municipio *i*. *i*=1,…, 125. Se tuvo en cuenta en el análisis la información de la variable X_i para los años 2001 a 2006.

Dado que los municipios tienen áreas disímiles entre si y con el fin de asegurar la comparabilidad de los valores de la variable en estudio en los diferentes municipios, se convirtió el número de homicidios en una tasa de homicidios por cada 100.000 habitantes de acuerdo con la siguiente expresión:

$$Tasa\ de\ Homicidios\ en\ municipio_i = \frac{x_i \times 100.000}{Población_i}$$

La información sobre la variable objeto de estudio fue procesada utilizando el programa GeoDa (Geodata Analysis Software), desarrollado por Luc Anselin y sus colaboradores en la Universidad de Illinois, para realizar Análisis Exploratorio de Datos Espaciales.

A continuación se presentan algunos de los reportes de los resultados que arroja el GeoDa y se realizan los análisis correspondientes.

7.2 Identificación de valores atípicos

7.2.1 Diagrama de Caja - Mapa

En la figura 1 se presentan los diagramas de Caja-Mapas para el año 2001. En este diagrama los municipios que aparecen resaltados en el mapa son aquellos que tuvieron las mayores tasas de homicidios y corresponden a los valores atípicos o *outliers*, a su vez esta información aparece en los puntos resaltados en la parte superior del diagrama de caja del lado derecho de la figura. En la parte inferior se observa la tabla donde aparecen los nombres de los municipios resaltados en el mapa y la subregión a la cual pertenecen.

Figura No. 1
Diagrama de Caja-Mapa para el año 2001

ıble : antioquia 09-11

MUNICIPI_1	NOMBRE	ZONA	POBL_2001	POBL_2002	POBL_2003	POBL_2004	POBL_2005	POBL_2006
86	YOLOMBO	Nordeste	19101	19326	19555	19788	20025	202
73	CONCEPCION	Oriente	5062	4915	4773	4638	4509	43
77	EL PENOL	Oriente	16243	16233	16229	16232	16241	162
84	SANTO DOMINGO	Nordeste	12310	12120	11933	11748	11567	113
78	GRANADA	Oriente	12077	11451	10862	10309	9789	93
81	SAN RAFAEL	Oriente	15415	14910	14428	13968	13530	131
108	COCORNA	Oriente	17410	16803	16219	15658	15119	146
80	SAN CARLOS	Oriente	18263	17614	16993	16397	15826	153
82	ALEJANDRIA	Oriente	4319	4186	4058	3935	3816	37
93	REMEDIOS	Nordeste	21353	21698	22049	22406	22769	231

Un análisis de los reportes permite afirmar que el fenómeno de violencia presenta una estructura en la cual el fenómeno aparece agrupado en municipios con altos índices de violencia. Los municipios

con mayores tasas de homicidios se encuentran en la misma zona para casi todos los años observados y algunos municipios se repiten en el transcurso de estos, esto nos lleva a conjeturar que el fenómeno de violencia no se distribuye en forma aleatoria.

7.2.2 Mapa de Caja o *Boxmap*

Figura No. 2
Boxmap para el año 2001

En la figura 2 se presentan un *BoxMa*p para uno de los años considerados. En este gráfico aparecen sombreados con cinco tonos de gris diferentes los municipios, de acuerdo con el valor que tome la variable en cada uno de los municipios. Este procedimiento estadístico permite identificar agrupaciones de municipios con características similares determinadas por el valor de la variable. En estos análisis reviste especial importancia los municipios con una tasa alta de homicidios y también se debe considerar como muy grave si además de tener una tasa alta de homicidios algunos de sus vecinos también tienen estas mismas características. Este gráfico permite identificar claramente que el fenómeno de la violencia no se produce de manera aislada sino que

existen grupos de municipios que a través de los años presentan altos índices de violencia como son los casos de San Carlos, San Francisco, Granada y Cocorná.

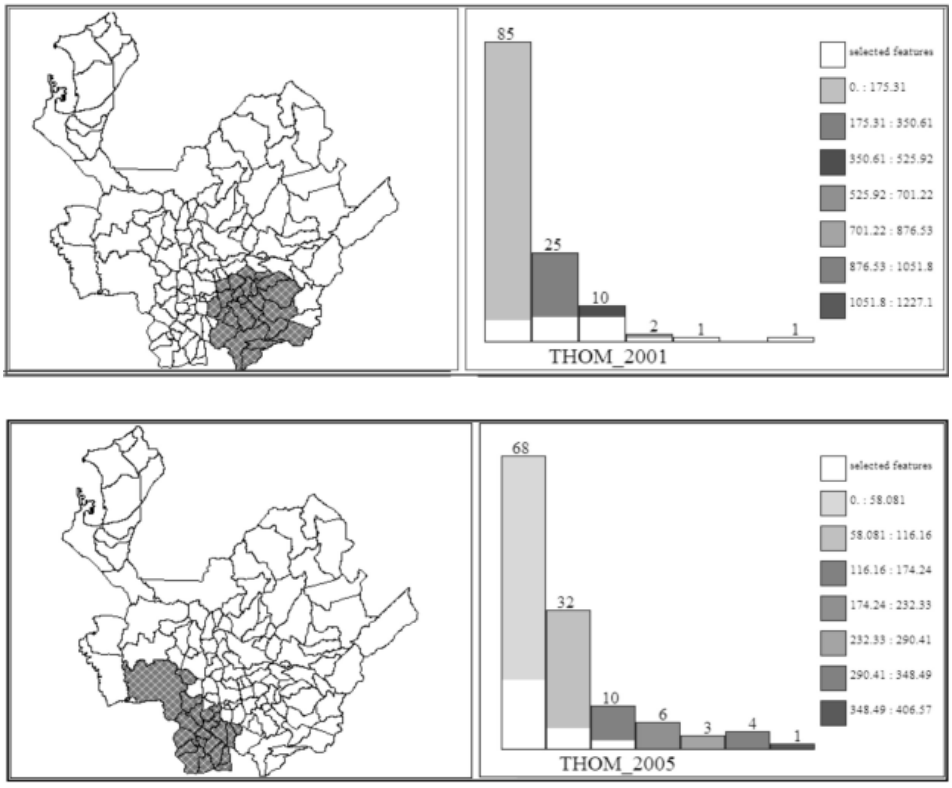
7.3 Identificación de la distribución de la variable

7.3.1 Histograma - Mapa

El histograma es la aproximación empírica a la distribución teórica de una variable. En este tipo de reporte utilizado es posible obtener también la aproximación a la distribución empírica de la variable en una subregión de la zona de estudio. Al comparar los dos histogramas, el de la zona de estudio y el de la subregión seleccionada es posible detectar la presencia de heterogeneidad espacial cuando el histograma regional no presenta una estructura similar al histograma general, esto significa que no hay uniformidad de los efectos espaciales entre la zona de estudio y la región.

En la figura 3 se presentan los histogramas de la variable Tasa de Homicidios para los años 2001 y 2005. De acuerdo con el reporte podemos afirmar que la heterogeneidad esta presente en la variable objeto de estudio ya que la distribución empírica en alguna de las subregiones difiere de la distribución empírica global. Esto se ilustra en el histograma para el año 2001 donde la distribución empírica de la subregión oriente, la cual se encuentra resaltada con blanco sobre el histograma general del departamento, no presenta una estructura similar a la del departamento. Esto se debe al hecho de que algunos de los municipios que conforman la subregión oriente, como son Concepción, El Peñol, Granada, San Rafael, Cocorna, San Carlos y Alejandría, tuvieron las mayores tasas de homicidios en dicho año, lo cual ocasiona que la variable en estudio no presente un comportamiento similar en la subregión analizada que en todo el departamento. Al analizar el gráfico correspondiente al año 2005, donde se resaltó la subregión suroeste se concluye que la subregión tiene una distribución empírica similar a la de la distribución del total de municipios, por lo cual se puede concluir que no hay heterogeneidad espacial entre la zona de estudio y la subregión.

Figura No. 3
Histograma para el año 2001 - Subregión Oriente resaltada
y año 2005 - Subregión Suroeste



7.3.2 Mapa de cuantiles

En la figura 4 se presentan el mapa de cuartiles para el año 2001 donde aparecen la distribución de las unidades espaciales de acuerdo con los cuartiles. Los municipios que pertenecen al mismo cuartil tienen el mismo color. La interpretación de este reporte es similar a la del BoxMap, ya que se puede ver como se agrupan los municipios que pertenecen a cada uno de los cuartiles, dentro de la zona de estudio.

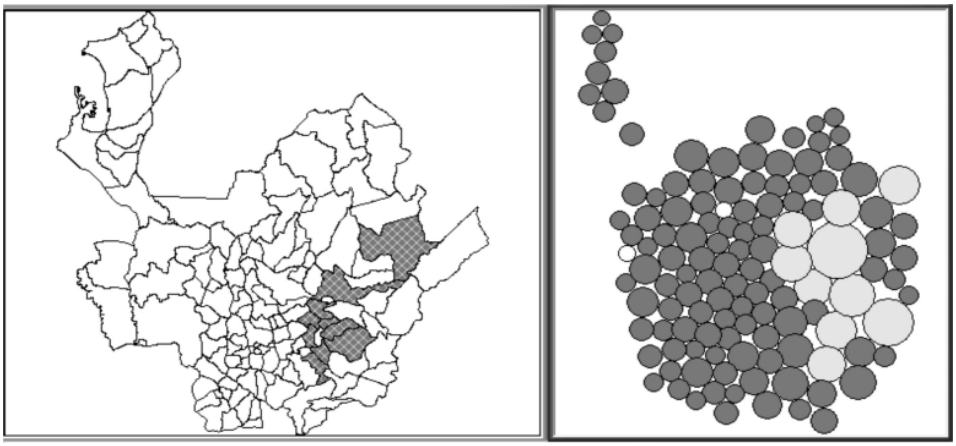
Figura No. 4
Mapa de Cuartiles para el año 2001

7.4 Identificación de efectos espaciales

7.4.1 Cartograma

La figura 5 corresponde al cartograma para el año 2001, donde cada unidad espacial es reemplazada por un círculo cuya área es proporcional al valor de la variable en dicha unidad espacial. En el cartograma las unidades espaciales con los valores atípicos se resaltan con colores diferentes del resto de las unidades, en este caso aparecen resaltados con color gris claro. El cartograma permite la identificación de las unidades con valores atípicos y adicionalmente permite comparar visualmente la relación que tienen las unidades con valores atípicos y las unidades con valores no atípicos.

Figura No. 5
Cartograma para el año 2001

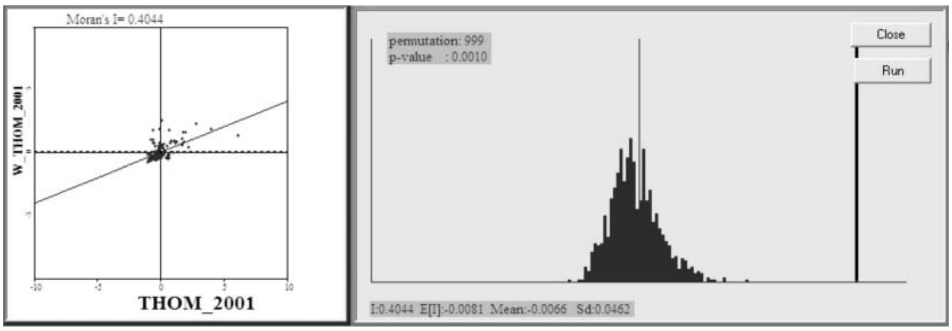


7.4.2 Diagrama de dispersión de Moran

El *scatterplot* de Moran es un diagrama de dispersión donde se representa la variable en estudio y el retardo espacial de dicha variable. El estadístico de prueba *I* de Moran para contrastar la autocorrelación espacial es el estimador de la pendiente de la regresión por mínimos cuadrados ordinarios. En la figura 6 se presentan el diagrama de dispersión de Moran del año 2001. En este tipo de reporte se observa en la parte superior del diagrama de dispersión el valor del estadístico de prueba *I* de Moran. En la parte derecha del diagrama aparece el calculo de la media de *I E(I)*, la desviación estándar *V(I)* y el valor *P* del estadístico de prueba. La conclusión para cada uno de los años considerados es el rechazo de la hipótesis nula de no autocorrelación espacial, es decir existe autocorrelación espacial entre los datos considerados, lo cual implica que estudios que utilicen modelos econométricos en los cuales en su especificación se tenga la variable tasa de homicidios, y en los cuales no se tenga en cuenta en la modelación la presencia de autocorrelación son de muy poco tienen valor científico, debido a un error de especificación. Este test nos indica que la distribución de los

homicidios en el Departamento no se distribuye de manera aleatoria lo cual constituye un punto de partida para el desarrollo de políticas que sobre la violencia diseñe el gobierno.

Figura No. 6
Diagrama de Dispersión de Moran para el año 2001



Conclusiones

Los resultados del caso de estudio sobre la violencia en Antioquia indican que este fenómeno no tiene una distribución aleatoria sino que presenta patrones espaciales de comportamiento, donde la violencia en un municipio contagia a los municipios vecinos. Por lo tanto las políticas gubernamentales deben estar diseñadas a controlar estos focos donde se concentra la mayor tasa de homicidios.

El anterior resultado confirma la necesidad de considerar el espacio como elemento estructural fundamental en el estudio de fenómenos sociales y económicos.

Siempre que se trabaje con variables georreferenciadas la primera etapa del análisis lo debe constituir el AEDE, con el fin de evaluar la presencia de efectos espaciales entre los datos, ya que de no realizarse este análisis el modelo tendrá errores de especificación.

Bibliografía

ANSELIN Luc, Spatial Econometrics: Methods and Models, Kluwer Academic Publishers, Dordrecht/Boston/London , 1988.

CASE Anne, ROSEN Harvey and HINES James. (1993). Budget Spillovers and Fiscal Policy Interdependence: Evidence from the States. Journal of Public Economics, 52(3): 285-307.

Centro de Análisis Político, Escuela de Ciencias y Humanidades Universidad EAFIT, Análisis de los Indicadores de Vida en el departamento de Antioquia 2001-2007, Medellín, 2007.

CHASCO Yrigoyen C., Tesis Doctoral: Econometría Espacial Aplicada a la Predicción -Extrapolación de Datos Microterritoriales, Consejería de Economía e Innovación Tecnológica, (2003).

FLINT, C., HARROWER M., & EDSALL, R., But How Does Place Matter? Using Bayesian Networks to Explore a Structural Definition of Place. Paper presented at the New Methodologies for the Social Sciences Conference. University of Colorado at Boulder, 2000.

GOODCHILD, M., A spatial analytical perspective on geographical information systems, International Journal of Geographical Information Systems, 1987, 1, 327-334.

HAINING R., Spatial Data Analysis, Theory and Practice. Cambridge, Cambridge University Press, UK, 2003.

MORAN P. A. P.,Notes on Continuous Stochastic Phenomena, Biometrika, Vol. 37, No. 1/2. (Jun., 1950), pp. 17-23.

MORENO R. y VAYA, E., Técnicas econométricas para el tratamiento de datos espaciales: La econometría espacial, Universitat de Barcelona, 2000, Barcelona

TOBLER, W. R., A computer movie simulating urban growth in the Detroit region. Economic Geography, 1970, 46(2),234-240.

VAYA E., LOPEZ-BAZO E., MORENO R. y SURIÑACH J., Economic growth and spatial externalities, Comunicación presentada en el 45 Noth American Meeting of the Regional Science Association International. Santa Fe, New Mexico, USA.

VILALTA y PERDOMO Carlos Javier, Cómo enseñar autocorrelación espacial, Economia, Sociedad y Territorio, Volumen v 18, 2005, pp. 323-333.