Cuadernos Latinoamericanos de Administración
ISSN: 1900-5016
cuaderlam@unbosque.edu.co
Universidad El Bosque
Colombia

de Corso Sicilia, Giuseppe Bernardo; Pinilla Rivera, Maribel; Gallego Navarro, Jaime
Métodos gráficos de análisis exploratorio de datos espaciales con variables espacialmente distribuidas
Cuadernos Latinoamericanos de Administración, vol. XIII, núm. 25, july-december, 2017, pp. 92-104
Universidad El Bosque
Bogotá, Colombia

Recibido: **27 | 04 | 2017** Aprobado: **13 | 12 | 2017**

Giuseppe Bernardo de Corso Sicilia [1] Maribel Pinilla Rivera [2] Jaime Gallego Navarro [3]

Artículo de Revisión

Métodos gráficos de análisis exploratorio de datos espaciales con variables espacialmente distribuidas

Methods Exploratory Analysis Graphs Spatial Data With Distributed Spatially Variable

JEL: M10; M14; M20.

[1]**PhD y Máster en análisis de políticas públicas** de la Universidad Simón Bolivar de Caracas (Venezuela). Economista de la University of Tampa Florida, EE.UU. Docente de la Universidad Jorge Tadeo Lozano. giuseppeb.decorsos@utadeo.edu.co

[2]**Administradora de Empresas.** Especialista en Gerencia Pública, Magister en Ciencias Económicas y candidata a Doctora en Modelado en Política y Gestión Pública de la Universidad Jorge Tadeo Lozano. Investigadora, docente de planta y directora del grupo de investigación en Estudios Ambientales GEA.UD, de la Universidad Distrital Francisco José de Caldas. mpinillar@udistrital.edu.co

[3]**Administrador Ambiental.** Grupo de investigación de Estudios Ambientales GEA.UD

Resumen. En el presente artículo se exponen, de manera compacta, los aspectos teóricos más relevantes de econometría espacial en la etapa de exploración estadística de datos, mediante las técnicas del análisis exploratorio de datos espaciales (AEDE), así como la fase previa a la formulación del modelo econométrico espacial. Con este artículo, se brindan herramientas para describir y visualizar distribuciones espaciales, que permitan darle validez a un modelo econométrico y que herramientas de la econometría tradicional no incorporan al rezagar efectos espaciales. La metodología utilizada se desarrolla a partir del AEDE y sus aplicaciones; para tal fin, se analizan las técnicas gráficas y estadísticas del AEDE que ofrece el software GeoDa 1.0.1, desarrollado por el profesor Luc Anselin de la Arizona State University, en dos dimensiones: (1) el análisis de datos univariante, es decir, se estudian las características de distribución espacial con respecto a una sola variable y (2) el análisis de datos multivariante, que involucra más de dos variables. Finalmente, se concluye que el AEDE debe constituir el primer eslabón en un análisis para la toma de decisiones en investigaciones de tipo ambiental, social y económica, cuyas técnicas principales a través de la estadística y la representación gráfica, posibilitan el análisis de las distribuciones espaciales y agrupamientos espaciales.

Palabras clave → análisis exploratorio de datos espaciales, distribución espacial, tendencia espacial, esquemas de asociación espacial.

Abstract. In this article, the most relevant theoretical aspects of spatial econometrics are presented in a compact manner in the stage of statistical data exploration using the techniques of exploratory spatial data analysis (AEDE), as well as the phase prior to the formulation of the model. econometric. This article provides tools for describing and visualizing spatial distributions that allow validation of an econometric model and that traditional econometric tools do not incorporate when delaying spatial effects. The methodology used is developed from the AEDE and its applications; for this purpose, we analyze the statistical and statistical techniques of the AEDE offered by GeoDa 1.0.1 software, developed by Professor

Luc Anselin, of the Arizona State Uninersity, in two dimensions: (1) the univariate data analysis; that is, the spatial distribution characteristics are studied with respect to a single variable and (2) the multivariate data analysis, which involves more than two variables. Finally, it is concluded that the AEDE should be the first link in an analysis for decision making in environmental, social and economic research whose main techniques are statistics and graphic representation, to enable the analysis of spatial distributions and spatial groupings.

Keywords → Exploratory Analysis Spatial Data, spatial distribution, spatial trend, spatial association schemes.

Introducción

En la actualidad no existen técnicas completas, que permitan una coherente descripción y visualización de distribuciones espaciales, que den validez a un modelo econométrico, herramientas que la econometría tradicional no incorpora al rezagar efectos espaciales. De esta necesidad, nace la pregunta de investigación de ¿porque son necesarios en los estudios sociales, ambientales y económicos, los análisis exploratorios de datos espaciales (AEDE)? Para entender un poco estas herramientas, es necesario acercarnos a una definición del análisis exploratorio de datos espaciales (AEDE), concebido como una disciplina dentro del análisis exploratorio de datos (AED).

Los métodos gráficos y visuales del Análisis Exploratorio de Datos, se usan para identificar las propiedades de los datos con el fin de detectar patrones en datos, formular hipótesis a partir de los datos y aspectos de la evaluación de modelos. El AEDE, puede plantearse desde el punto de vista de la Geo-estadística o por la econometría espacial, donde la Geo-estadística es una rama de la Estadística que trata fenómenos espaciales (Journel & Huijbregs, 1978), cuyo interés principal es la estimación, predicción y simulación de dichos fenómenos (Warrick & Myers, 1987). Al respecto, se reconoce como una rama de la Estadística tradicional que, parte de la observación de que la variabilidad o continuidad espacial de las variables distribuidas en el espacio tienen una estructura particular, que se estudia mediante las dependencias entre ellas. De otro lado, la Econometría espacial se ocupa de la dependencia espacial y la heterogeneidad espacial, aspectos críticos de los datos utilizados por los científicos regionales.

Por lo anterior, en este artículo se presenta la aplicación de las principales técnicas del AEDE, combinadas con el análisis estadístico gráfico. Esto hace posible, el estudio de las distribuciones espaciales y sus valores atípicos, esquemas de asociación espacial y agrupamientos espaciales. Para ello, se utiliza el programa que ha sido desarrollado por el profesor Luc Anselin, de la Arizona State Uninersity, con el cual se presenta la capacidad y las posibilidades del AEDE. La versión más reciente del programa, está disponible en internet (http://sal.agecon.uiuc.edu/geoda_main.php) y es de acceso libre.

Materiales y métodos

El análisis exploratorio de datos espaciales (AEDE), se define como el grupo de técnicas que describen y visualizan las distribuciones espaciales, identifican localizaciones atípicas, descubren esquemas de asociación (auto-correlación espacial) y sugieren estructuras en el espacio geográfico (heterogeneidad espacial) (Ver Hoef, 1993); por consiguiente, el AEDE es más una técnica descriptiva (estadística) que confirmatoria (econométrica) (Chasco Yrigoyen, 2003). Al respecto, se reafirma que el análisis exploratorio de datos, es el estudio previo al análisis confirmatorio de datos espaciales, en los que se formulan modelos de regresión y se realiza la estimación de parámetros muestrales.

Uno de los componentes más relevantes dentro del AEDE e incluso dentro del AED, es el análisis gráfico. Este, combinado con técnicas de análisis estadístico, da origen a lo que suele denominarse visualización científica (Smouse, Long & Sokal, 1986), la cual permite extraer toda la información posible y de manera eficiente, cuando se trabaja con grandes bases de datos; simultáneamente, genera técnicas gráficas, con la capacidad de trabajar con la totalidad de las observaciones o, si se desea, analizar parcialmente un determinado conjunto de datos para establecer comportamientos, tendencias, puntos atípicos, entre otros.

En este sentido, un método eficiente de visualización científica del AEDE, es aquel que permite identificar dos características básicas de las distribuciones espaciales: suavizado (smooth) y asperezas (rought) (Velleman, 1981; Phillips, 1985).

El suavizado (smooth), que, en el contexto temporal, es la tendencia central de la variable determinada mediante un elemento central como la mediana y medidas de dispersión, permite determinar tendencias y patrones de asociación espacial en un esquema global de análisis, es decir, logra identificar patrones de asociación espacial representados mediante la autocorrelación espacial global. Por su parte, las asperezas (rought), son un análisis local que identifica la presencia de puntos atípicos (outliers) en distribuciones espaciales.

Dentro de las investigaciones sociales, ambientales, económicas y del territorio, el análisis del espacio y la localización, han sido variables inquietantes para la toma de decisiones en los sectores públicos y privados, y la puesta en marcha de

planes de manejo territorial, por ejemplo. Por ese motivo, este artículo brinda técnicas de análisis exploratorio de datos espaciales, soportados en técnicas gráficas, que permitan un mejor entendimiento de las problemáticas existentes.

Las principales técnicas expuestas en este artículo de análisis exploratorio de datos reticular, incluidas en el programa GeoDa, son AED general, cuyas tendencias espaciales son visualizadas mediante: (i) el histograma de frecuencias, (ii) el diagrama de dispersión, (iii) el gráfico de coordenadas paralelas y (iv) el gráfico de dispersión en 3D; los atípicos espaciales, se observan a través de los diagramas de caja. Por otra parte, el AEDE reticular, analiza la tendencia espacial mediante: (i) mapas temáticos, (ii) mapa dinámico, (iii) gráficos condicionales, (iv) diagrama de dispersión de Moran y (v) diagrama de dispersión de Moran multivariante. Los respectivos atípicos espaciales, se analizan con: (i) mapa de caja, (ii) mapa de percentiles, (iii) cartograma, (iv) gráficos LISA y (v) gráficos LISA multivariantes.

Resultados

En los últimos años, se han propuesto gran cantidad de métodos gráficos para el análisis exploratorio de datos (AEDE), aunque pocos le den el interés y la efectividad de todos ellos (Haining, 2000; Wagner, 2003). Se podría afirmar, entonces, que el método gráfico de AEDE proporciona resúmenes rápidos y visuales de las características de datos esenciales, herramienta de gran ayuda para la toma de decisiones.

Análisis univariante espacial

Se presentan las técnicas de visualización gráfica, más usadas en el contexto de la econometría espacial para el análisis espacial univariante, las cuales permiten identificar tendencias, esquemas de asociación y distribución espacial. Para ilustrar los ejemplos, se trabaja con base en datos correspondientes a variables ambientales de Colombia, que están asociadas a la temperatura media, nivel de pluviosidad y nivel de humedad relativa, en todos los departamentos de Colombia para el año 2015.

Representación de la tendencia central.

A menudo, cuando se describen diferentes grupos de observaciones, es necesario resumir la información en un conjunto de datos menos extenso que los originales y que permita identificar el comportamiento de los datos, con base en criterios de categorización cuantitativa. Dentro de este grupo se cuenta con: mapas de cuantiles, histogramas de frecuencia y mapas de desviación típica.

Mapas temáticos (cuantiles)

Los mapas temáticos, en general, consisten en la representación cartográfica de una variable geográfica. Esta representación de la variable, en un mapa, puede llevarse a cabo mediante símbolos y colores, que pongan de manifiesto el valor de una variable en cada una de las unidades geográficas consideradas (países, regiones, etc.) (Anselin, 1995). Puede utilizarse un color/símbolo diferente para cada valor o para cada intervalo de valores de la variable. Dentro del AEDE reticular, los mapas temáticos más importantes para la representación de la tendencia espacial de una variable, son el mapa de cuantiles y el mapa de la desviación típica (Chasco Yrigoyen, 2003).

Los mapas temáticos, son representaciones cartográficas que identifican fenómenos geográficos, como

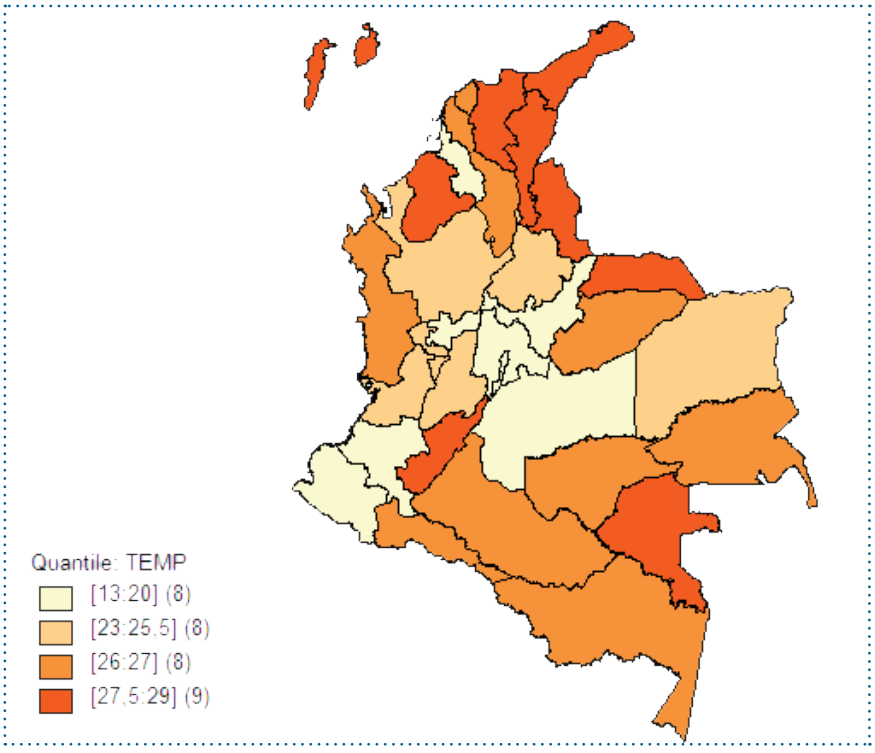


Figura 1. *Mapa de cuantiles para la temperatura promedio por departamentos en Colombia, 2015. Fuente: elaboración propia a partir de GeoDa 1.0.1., 2016.*

distribución, densidad o relación de datos de una variable espacialmente distribuida, mediante el uso de recursos visuales: colores, símbolos o cualquier forma que ponga de manifiesto la disconformidad de valores en una misma variable. Dentro de los mapas temáticos, se cuenta con los mapas de cuantiles, que representan el comportamiento de una variable espacialmente distribuida, para lo cual, los datos se dividen y se agrupan en una serie de categorías denominadas cuantiles, que en lo posible contienen igual número de observaciones. Cuando las categorías se dividen en 4, 5, 6, se les denomina, respectivamente: cuartiles, quintiles, sextiles y así sucesivamente.

Como se observa en la figura 1, en el mapa de cuantiles para la temperatura promedio en Colombia (con datos de 2015), existe una tendencia en la región Caribe, de valores altos de la temperatura (27,5 °C-29 °C) con respecto a los departamentos de La Guajira, Magdalena, Norte de Santander, Cesar y Córdoba; asimismo, se evidencia otra tendencia en la región amazónica, de valores medios de temperatura (23 °C-27 °C) que corresponden a los departamentos de Amazonas, Putumayo, Guaviare y Guainía. Es decir, la temperatura promedio tiene una distribución espacial con patrones de asociación, que dependen de la ubicación geográfica; sin embargo, existen valores que son muy parecidos, pero se definen en cuantiles diferentes. Por ejemplo, el departamento de Vaupés, tiene una temperatura promedio de 27,5 °C; no obstante, es clasificado dentro de los cuantiles con el rango más alto.

Histograma de frecuencias

Este permite visualizar la distribución espacial de datos de naturaleza continua. El histograma se compone de una serie de barras gráficas, en las que la altura de cada una corresponde a la frecuencia de los valores representados. En el eje de las abscisas, se representan los valores de la variable, dividida en intervalos. Por su parte, en el eje de las coordenadas, se expresan las frecuencias absolutas de cada intervalo.

El histograma de frecuencias (figura 2), es un gráfico estadístico clásico en el AED. El programa GeoDa, calcula histogramas de frecuencias de las variables geográficas para distintas clasificaciones, aunque el número por defecto es 7. Cada una de las barras del histograma, tiene un color y es posible hacer una selección en el histograma para ver sobre el mapa las observaciones a las que corresponde (Unwin, 2000; Leduc, Drapeau, Bergeron & Legende, 1992).

Los histogramas son útiles, cuando la variable en cuestión presenta valores muy parecidos, donde los mapas de cuantiles no pueden definir estos, porque no se puede asignar un número igual de observaciones a los diferentes grupos.

Mapa de desviación típica

Este mapa (figura 3) agrupa las observaciones que, según sus valores, caigan dentro de un rango estandarizado, es decir, como un número determinado de unidades de la desviación típica a partir de la media. Las categorías en las que se divide la variable, se corresponden con múltiplos de la desviación típica de la variable. O sea, clasifica las observaciones en un rango estandarizado de valores a partir de la media (Casetti & Poon, 1995; Warrick & Myers, 1987).

Los valores con mayor grado de dispersión de valores bajos con respecto al valor de la temperatura promedio, son los correspondientes a Bogotá y Nariño. Nótese que estos valores, no pueden detectarse en un mapa de cuantiles.

Representación de puntos atípicos

Antes de entrar con detalle en el análisis gráfico de los puntos atípi-

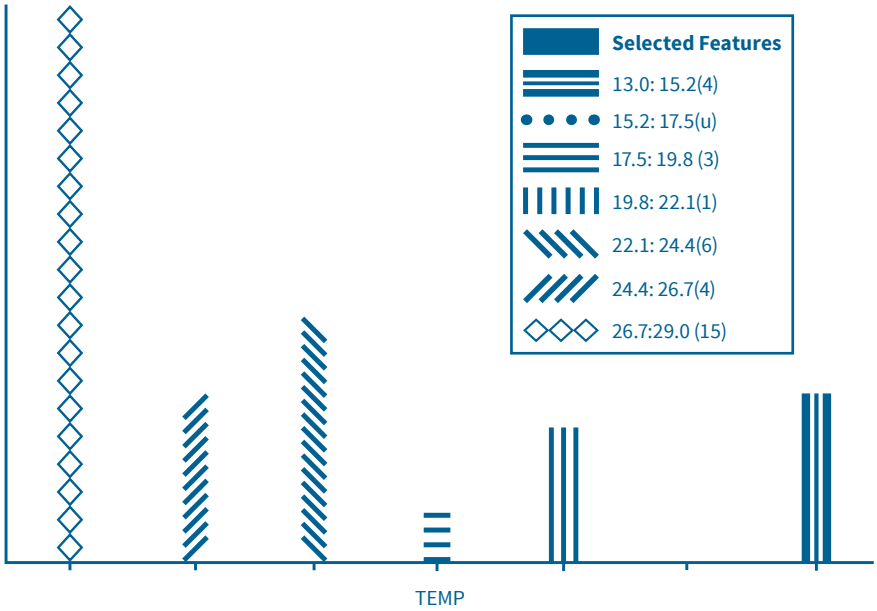


Figura 2. *Histograma de frecuencias para la temperatura promedio por departamentos en Colombia, periodo 2015. Fuente: elaboración propia a partir de GeoDa 1.0.1., 2016.*

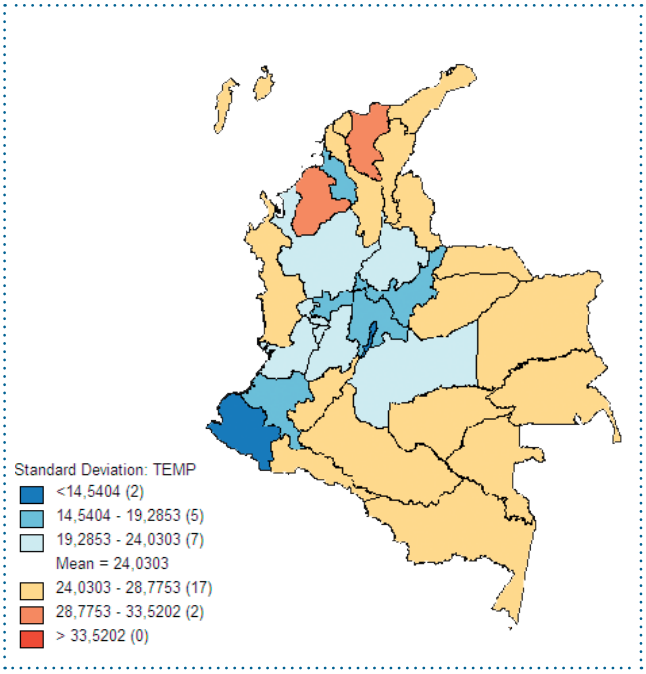


Figura 3. *Mapa de desviaciones típicas para la temperatura promedio, periodo de 2015, por departamentos en Colombia. Fuente: elaboración propia a partir de GeoDa 1.0.1., 2016.*

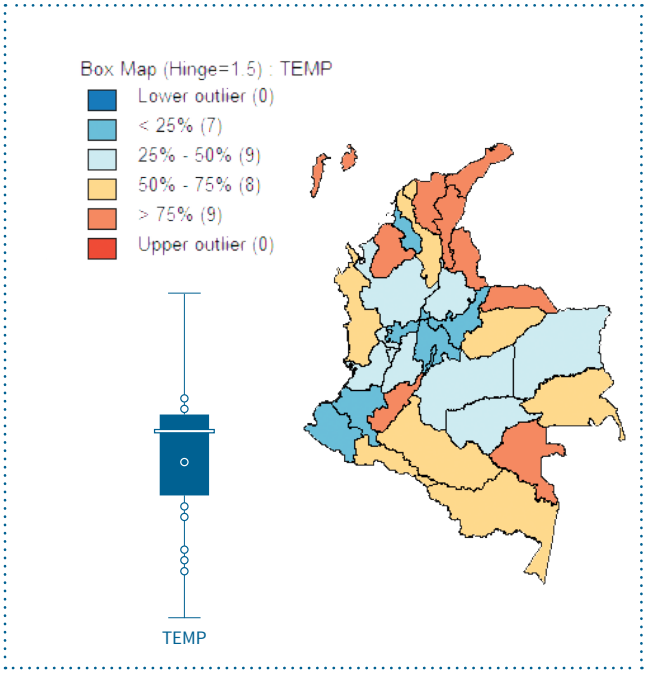


Figura 4. *Box Plot para la temperatura promedio, periodo de 2015, por departamentos en Colombia. Fuente: elaboración propia a partir de GeoDa 1.0.1., 2016.*

cos, es necesario precisar que un punto atípico, tiene que ver con aquellos elementos discontinuos en una variable, es decir, representan valores excesivamente bajos o altos que, generalmente, no son significativos y tienden a distorsionar el comportamiento de la variable. Los puntos atípicos (elementos de discontinuidad en una variable) son valores de la variable excepcionalmente bajos/altos, que pueden no ser representativos en la distribución general y afectarían el comportamiento de los contrastes estadísticos (Acevedo & Velásquez, 2008; Cressie, 2001).

Dentro de los análisis del AED, la presencia de puntos atípicos implica la existencia de errores de medida, que expresan situaciones extrañas en el comportamiento de los datos y no aportan información relevante; por lo cual, en algunos casos, se aconseja eliminarlos; pero, si la cantidad de datos atípicos es relevante y eliminarlos implica perder la estructura del modelo, es necesario usar transformaciones para suavizar las bases de datos (Moreno & Vayá, 2000).

Diagrama de caja y bigotes (Box Plot)

El diagrama de caja y bigotes (figura 4), es un gráfico representativo de las distribuciones de un conjunto de datos, en cuya construcción se usan cinco medidas descriptivas de estos, a saber: mediana, primer cuartil, tercer cuartil, valor máximo y valor mínimo (Duncan, 1991; Okabe, Satoh & Sugihara, 2009). Esta presentación visual, asocia las cinco medidas que suelen trabajarse de forma individual. Al mismo tiempo, presenta información sobre la tendencia central, dispersión y simetría de los datos de estudio. Además, permite identificar, con claridad y de forma individual, observaciones que se alejan de manera poco usual del resto de los datos. A estas observaciones, se les conoce como valores atípicos (Graham & Glaister, 2003).

Por su facilidad de construcción e interpretación, también ayuda a comparar a la vez varios grupos de datos, sin perder información ni saturarse de ella. Esto ha sido particularmente importante, en el momento de escoger esta representación para mostrar la opinión de los estudiantes respecto de la actuación docente, mediante las diversas preguntas del instrumento utilizado (Whittle, 1954; Vilalta Perdomo, 2005).

La construcción del rectángulo, implica el cálculo del primer cuartil (en el que se ubica máximo el 25 % de los datos) y el tercer cuartil (donde se pone máximo el 75 % de los datos), así como de la mediana (que corresponde al valor del segundo cuartil). El cálculo de los límites inferior y superior se obtiene restando y sumando respectivamente, a la mediana el producto de los valores del tercer (primer) cuartil por 1,5

veces (o 3 veces, dentro de un criterio un poco más estricto, si se desea identificar puntos atípicos con mayor exactitud) el recorrido intercuartílico.

El mapa de Box Plot, indica que existen datos con un grado moderado de dispersión por debajo de la mediana, entre el primer cuartil y el límite inferior en el Box Plot (están representados en el mapa con color azul), y por encima de la mediana entre el tercer cuartil y el límite superior (marcados con color rojo). Sin embargo, el Box Plot no identifica valores atípicos representativos.

Cartograma.

Los cartogramas (figura 5), son una forma de representación de puntos atípicos dentro de las ubicaciones correspondientes en el mapa; asimismo, la diferencia o discontinuidad de valores se muestra, en la proporción del tamaño de la observación con respecto a las demás. El cartograma, permite la identificación de las unidades con valores atípicos y, adicionalmente, facilita la comparación visual de la relación que tienen las unidades con valores atípicos y las unidades con valores no atípicos (Wartenberg, 1985; Epperson & Li, 1996).

En el mapa de la desviación estándar para el nivel de pluviosidad, se evidencia la existencia de un valor atípico que, en el Box Plot, se ubica por encima de la mediana, situándose arriba del límite superior. En el cartograma, el círculo de color rojo, que, a su vez, es el más grande, representa este valor, correspondiente al nivel de pluviosidad más alto, el cual se presenta en el departamento del Choco que, para el año 2015, fue de 9000 m. m.

Análisis multivariante espacial

El análisis multivariante (AM), es la parte de la estadística y del análisis

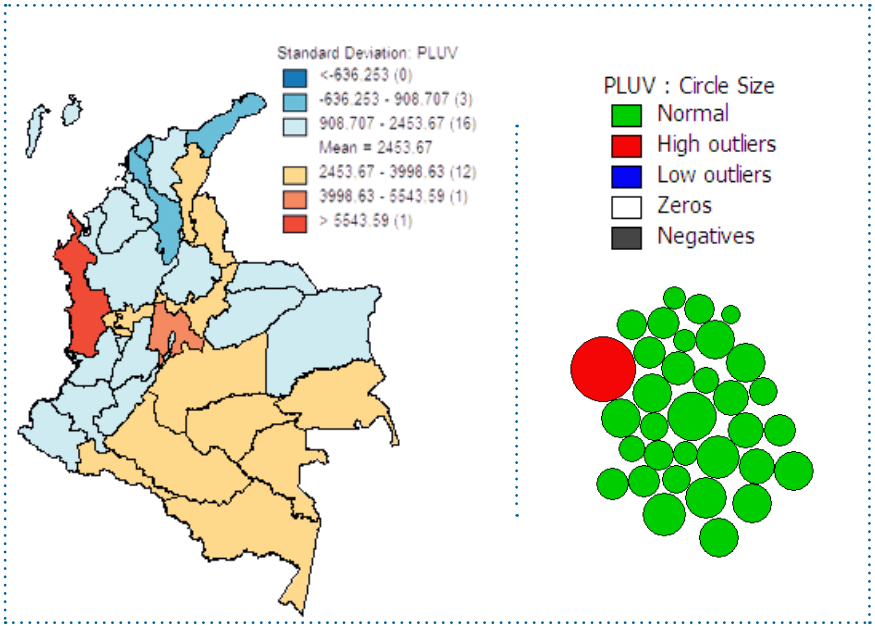


Figura 5. *Cartograma de pluviosidad promedio, periodo de 2015, por departamentos en Colombia. Fuente: elaboración propia a partir de GeoDa 1.0.1., 2016.*

de datos que estudia, analiza, representa e interpreta los datos que resultan de observar más de una variable estadística, sobre una muestra de individuos. Las variables observables son homogéneas y correlacionadas, sin que alguna predomine sobre las demás. La información estadística en AM, es de carácter multidimensional; por lo tanto, la geometría, el cálculo matricial y las distribuciones multivariante desempeñan un papel fundamental (Sampson, 1987; Delfiner, 1979).

Los análisis multivariante, permiten analizar la relación entre múltiples variables en simultánea; es decir, identifican la incidencia de un grupo de variables independientes X, con respecto a una sola variable dependiente Y, medidas mediante un conjunto de observaciones o datos.

Diagramas de dispersión.

Los diagramas de dispersión, son nubes de puntos que representan la relación entre dos variables X y Y, o sea, permiten identificar el grado de correlación o dependencia entre variables. Las relaciones que pueden llegar a presentarse son de tipo lineal, cuando la intersección de cada dato de X con su respectivo dato en Y, forma una elipse en cualquier sentido; nula, cuando la distribución de los puntos no tiene una forma estructurada (por ejemplo, un círculo), y no lineal, si los puntos adoptan cualquier otra forma (logarítmica, exponencial, cuadrática, etc.). Pero, la visualización gráfica no es suficiente para determinar algún tipo de relación. Por tal motivo, se hace necesario el uso de un coeficiente que permita medir el grado de dependencia entre variables, como es el caso del coeficiente de correlación lineal, que representa el comportamiento de una variable dependiente Y, con respecto a una variable independiente X. Al respecto, el coeficiente de correlación lineal (r) se mueve en un rango de -1 a 1, en el que valores cercanos a 1 indican una relación lineal positiva directamente proporcional; valores cercanos a 0 evidencian que no existe un esquema de correlación definido entre variables y valores próximos a -1 indican la exis-

tencia de una relación lineal negativa, inversamente proporcional de X con respecto a Y.

Diagrama de dispersión espacio-temporal

El diagrama de dispersión espacio-temporal (figura 6), arroja el valor del coeficiente de correlación lineal r, denotado por la pendiente, asimismo, permite hacer análisis parciales de datos. Algunas observaciones sobre el coeficiente de correlación r para tener en cuenta son:

- *I r I = 1 Relación lineal perfecta*
- *I r I > 0,8 Relación lineal fuerte*
- *0,5 ≤ I r I ≤ 0,8 Relación lineal moderada*
- *0 ≤ I r I ≤ 0,5 Relación lineal débil*
- *I r I = 0 No existe una relación lineal*

Como se observa en el siguiente diagrama, el valor y signo del coeficiente denotado por la pendiente, indica que existe una relación lineal negativa muy débil (-0.0191), es decir, el grado de asociación lineal entre la temperatura y el nivel de pluviosidad es casi nulo. Por otra parte, como se observa en el diagrama, está divido en cuatro cuadrantes: los cuadrantes superior derecho e inferior izquierdo, representan esquemas de correlación positiva débil y correlación positiva fuerte, respectivamente; en consecuencia, los cuadrantes superior izquierdo e inferior derecho, muestran situaciones de correlación negativa o inversa débil y correlación negativa fuerte, respectivamente.

Diagrama de coordenadas paralelas

El diagrama de coordenadas paralelas, permite efectuar análisis con más de dos variables. El objetivo principal, es determinar grupos de datos con valores similares en las variables objeto de estudio y analizar el comportamiento de cada dato con respecto a cada una de las variables.

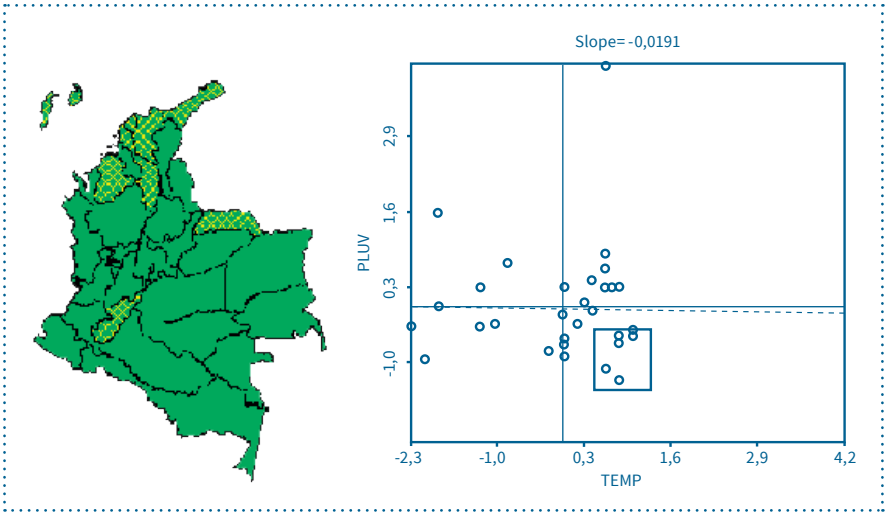


Figura 6. *Diagrama de dispersión espacio-temporal de la temperatura, con respecto al nivel de pluviosidad, en el periodo de 2015, por departamentos en Colombia. Fuente: elaboración propia a partir de GeoDa 1.0.1., 2016.*

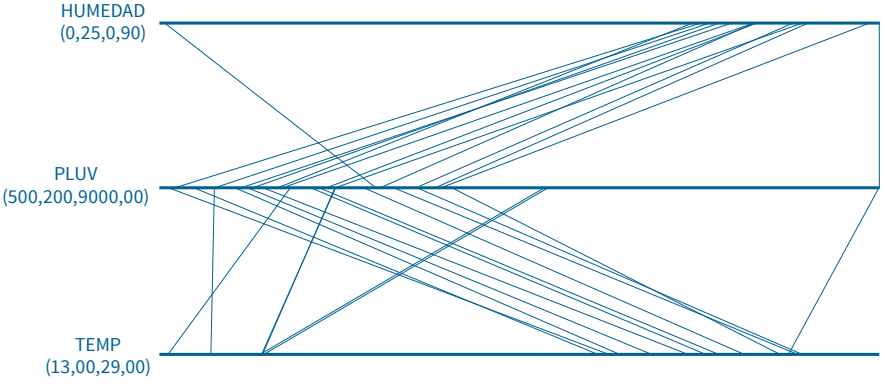


Figura 7. *Gráfico de coordenadas paralelas de la temperatura, el nivel de pluviosidad y la humedad relativa, en el periodo de 2015, por departamentos en Colombia. Fuente: elaboración propia a partir de GeoDa 1.0.1., 2016.*

En el eje Y, se representa cada variable mediante un segmento. El orden de las variables no es significativo en la interpretación estadística; entre tanto, en el eje X, se indican los valores de las variables: desde los inferiores en la izquierda del eje, hasta los superiores a la derecha de este. Cada variable es re-escalada, de forma que el valor mínimo se encuentre en el extremo izquierdo y el máximo en el extremo derecho.

En la Figura 7, se ilustra el comportamiento del departamento del Putumayo, que comienza con un valor alto para la humedad relativa (80 %); posteriormente, se desplaza hacia la izquierda, con valor medio más bajo del nivel de pluviosidad en proporción con la humedad relativa (3900 mm); finalmente, avanza hacia la derecha del eje hasta un valor alto de la temperatura (27 °C), ubicado en el tercer cuartil. Este tipo de análisis, es realmente eficiente cuando se manejan variables con la misma unidad de medida. La principal utilidad de este gráfico, consiste en la identificación de agrupamientos de valores en ciertas observaciones, que también pueden ser de naturaleza espacial.

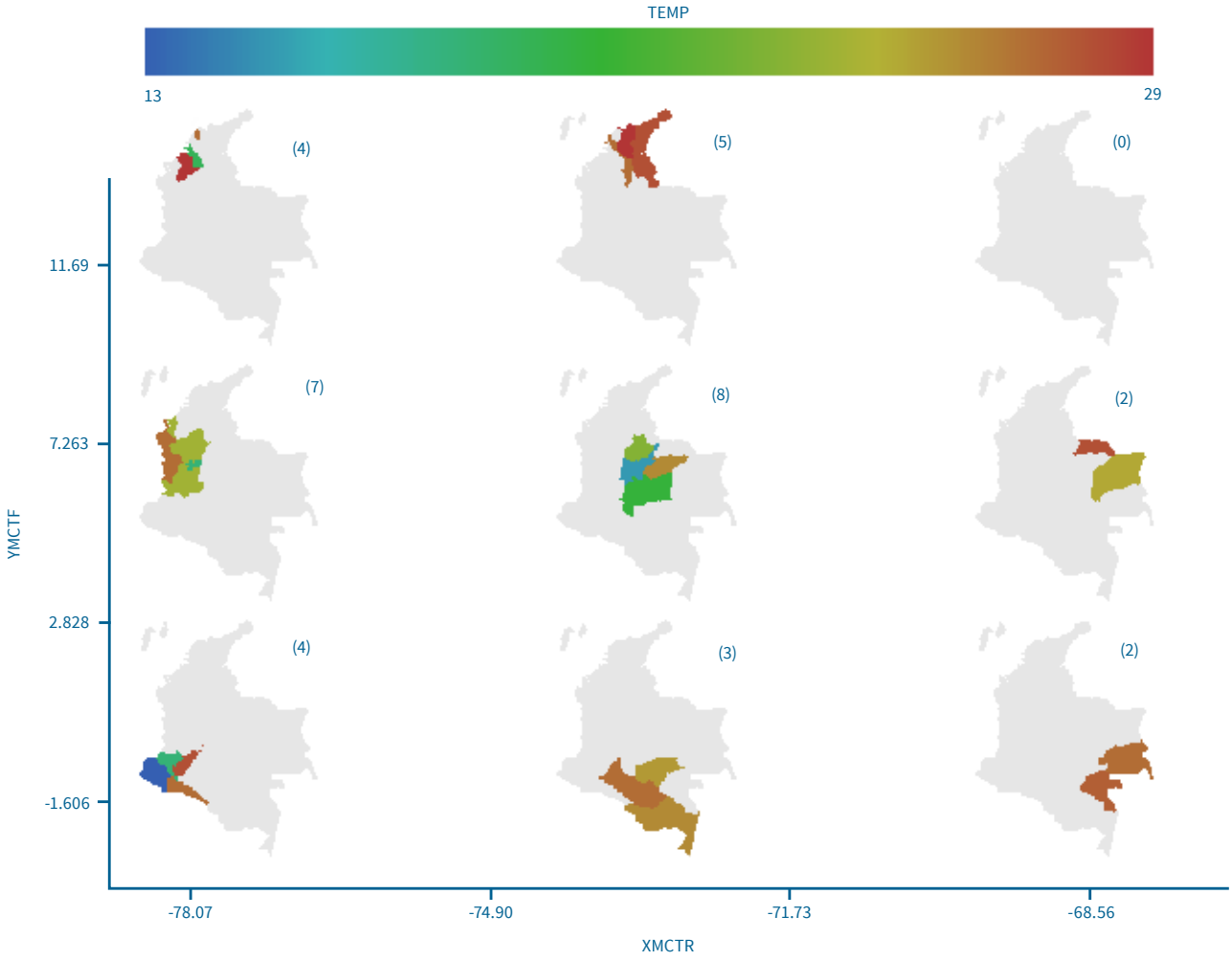


Figura 8. *Grafico condicional del NBI, por departamentos en Colombia, datos de 2015. Fuente: elaboración propia a partir de GeoDa 1.0.1., 2016.*

Gráficos condicionales

Los gráficos condicionales (figura 8), permiten analizar un conjunto de datos condicionados por dos variables que dividen la muestra en varias categorías. En los ejes, se representa cada una de las dos variables condicionadas (en general se trabaja con coordenadas espaciales X y Y) y se segmentan los ejes, respectivamente, en tres grupos de valores: altos, medios y bajos. Esto, da un total de nueve cuadrantes que están representados por mapas. Una tercera variable, se evalúa en función de las observaciones que estén comprendidas dentro de cada categoría.

Técnicas avanzadas del AEDE

Una vez analizados los métodos gráficos, que determinan la tendencia espacial de los datos, es posible plantear las hipótesis para contrastar, en relación con la dependencia espacial y generar las herramientas que permitan realizar los procesos de verificación estadística.

Contrastes de dependencia espacial global univariante.

Con el uso del test de dependencia espacial global univariante, se pueden identificar tendencias de asociación espacial de una variable en todo un espacio geográfico determinado.

Matrices de interacciones espaciales

La estructura espacial, suele expresarse formalmente mediante una matriz de interacciones espaciales. Esta también es llamada "matriz de pesos, ponderaciones, distancias o contactos espaciales". En esta matriz, cada unidad espacial se representa a la vez mediante una fila y una columna. En cada fila, los elementos no nulos de las columnas se corresponden con las unidades espaciales con-

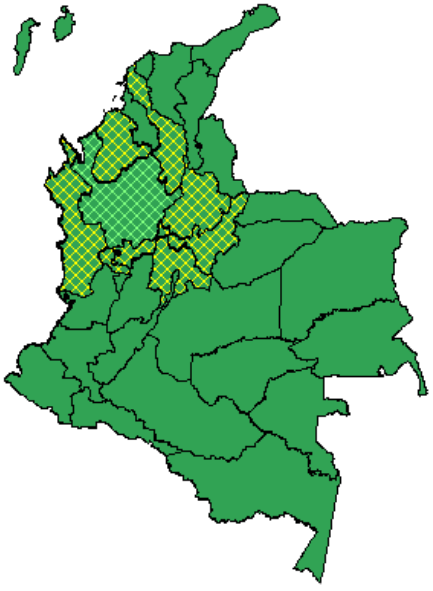
tiguas (Vayá, Moreno & Suriñach, 2002; Bradshaw, & Spies, 1992).

Los contrastes de dependencia o auto-correlación espacial, pueden basarse en una noción de contigüidad binaria entre las unidades espaciales. De acuerdo con este concepto, una situación de vecindad entre dos unidades espaciales, se podría expresar mediante valores de tipo 0-1 (Ord, 1975).

Esta definición de contigüidad requiere de la existencia de un mapa, a partir del cual se puedan obtener las fronteras entre unidades espaciales, que se creen de forma automática. Cuando dichas unidades se posicionan de forma irregular, fácilmente se pueden conocer las fronteras entre las distintas unidades geográficas; sin embargo, cuando las unidades pertenecen a una cuadrícula regular, la determinación de la contigüidad no es única (Dykes, 1998; Dale, 1999).

1	Amazonas	18	Guaviare
2	Antioquia	19	Huila
3	Arauca	20	Magdalena
4	Atlantico	21	Meta
5	Bogota	22	Nariño
6	Bolivar	23	Norte de Santander
7	Boyaca	24	Putumayo
8	Caldas	25	Quindio
9	Caqueta	26	Risaralda
10	Casanare	27	San Andrés y Providencia
11	Cauca	28	Santander
12	Cesar	29	Sucre
13	Choco	30	Tolíma
14	Cordoba	31	Valle del Cuaca
15	Cundina-marca	32	Vaupes
16	Guajira	33	Vichada
17	Guania		

Figura 9. *Mapa de convenciones de Colombia por departamentos. Fuente: elaboración propia a partir de GeoDa 1.0.1., 2016.*



DEPT	# Obs	Observaciones	DEPT	# Obs	Observaciones
1	3	32249	18	5	322117339
2	8	2615141386728	19	6	3021951511
3	3	10733	20	4	126164
4	2	206	21	7	331891019155
5	3	191521	22	2	2411
6	7	292014122824	23	3	28712
7	6	2824151032	24	4	192211
8	4	3026152	25	3	312630
9	7	2421191132118	26	6	3182530132
10	5	21153337	27	0	0
11	6	31229192430	28	5	7122362
12	5	282320166	29	2	614
13	3	31262	30	7	3115819262511
14	3	2962	31	5	2526301113
15	8	301957102182	32	4	117189
16	2	2012	33	5	172118103
17	3	333218			

Figura 10. *Mapa de contigüidades por departamento. Fuente: Elaboración propia a partir de GeoDa 1.0.1., 2016.*

Un primer paso para la formulación de las pruebas de contraste espacial de carácter univariante, está en estructurar los posibles sistemas de interacciones, basados en el uso de matrices. Precisamente, ahí se determina el mejor parámetro que identifica el esquema de relaciones entre cada variable, con la utilización de los criterios de contigüidad espacial. Para comprender mejor la utilidad mediante los criterios, considérese el siguiente mapa de convenciones: (figura 9).

A continuación, se muestran los resultados obtenidos mediante el criterio de contigüidad de la reina (criterio de contigüidad para ocho observaciones, donde serán vecinas de i las regiones que comparten algún lado o vértice con i) para identificar la vecindad por departamentos. Ver figura 10.

El departamento de Antioquia tiene ocho departamentos vecinos, según el criterio de la reina: Bolívar, Boyacá, Caldas, Chocó, Córdoba, Cundinamarca,Risaralda y Santander.

Test I de Moran

La herramienta auto-correlación espacial (I de Moran global) mide, simultáneamente, la auto-correlación espacial basada en las ubicaciones y los valores de las entidades. Dado un conjunto de entidades y un atributo asociado, evalúa si el patrón expresado está agrupado, disperso o es aleatorio. La herramienta calcula el valor del índice I de Moran y una puntuación z. El estadístico de prueba I de Moran para contrastar la auto-correlación espacial es el estimador de la pendiente de la regresión por mínimos cuadrados ordinarios (Bellehumeur & Legendre, 1998; Biondi, Myers & Avery, 1994).

Para construir el diagrama de dispersión de Moran de una variable específica, es necesario rezagar

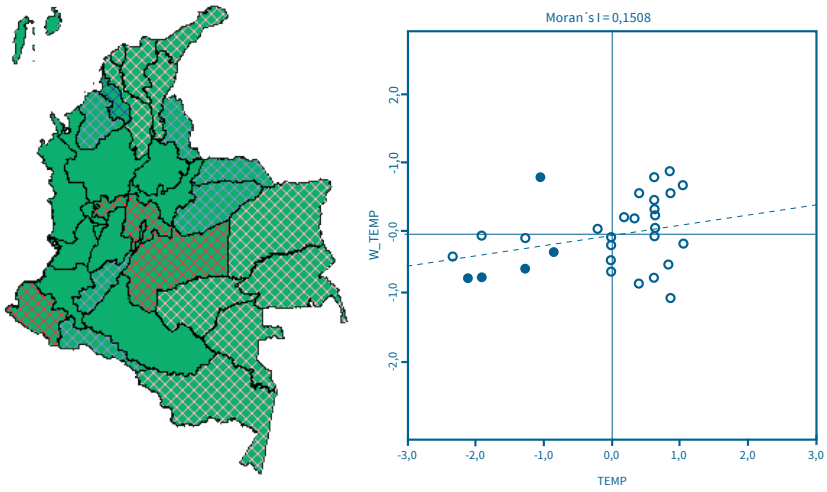


Figura 11. *Diagrama de dispersión I de Moran de la temperatura, en el periodo de 2015, por departamentos en Colombia. Fuente: elaboración propia a partir de GeoDa 1.0.1., 2016.*

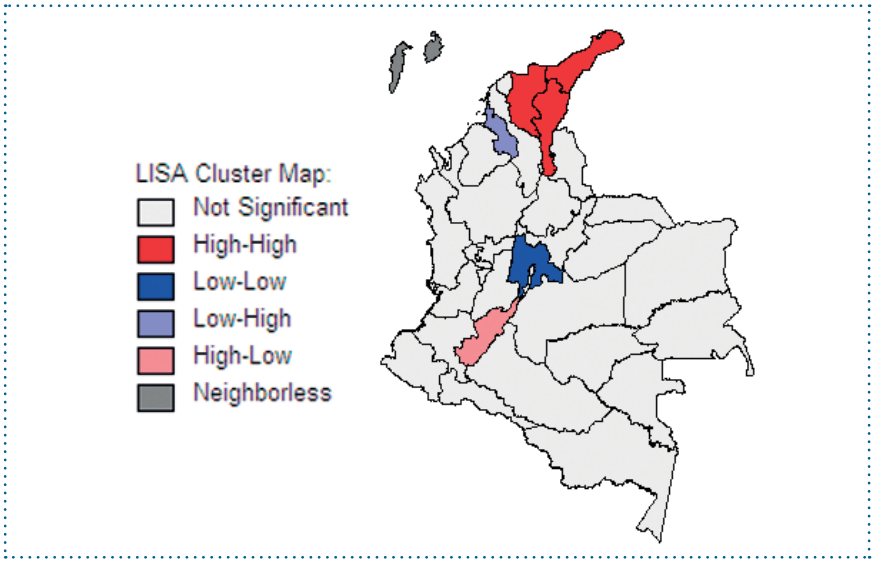


Figura 12. *Mapa LISA para la temperatura, en el periodo de 2015, por departamentos en Colombia. Fuente: elaboración propia a partir de GeoDa 1.0.1., 2016.*

espacialmente la variable en cada observación; este proceso, consiste en calcular un parámetro w y multiplicarlo por una observación i de la variable en cuestión, donde el parámetro w se obtiene al promediar los valores de la variable vecinos a i, en el orden de contigüidad especificado. Quedando así Wi:

$$Wi=\frac{Yi*Wi}{(Yi+Yk+Yl+\ldots+Yn)/n}$$

Donde Yi Yk Yl …Yn son las zonas contiguas a la región Yi. Los valores rezagados de la variable se ubican en el eje Y y los valores normales de la variable se ponen en el eje X. A continuación, se presenta el diagrama de dispersión para la variable temperatura.

El valor del estadístico I de Moran, sugiere que existe un esquema débil de auto-correlación espacial global positiva. Es decir, a partir del examen del diagrama de dispersión I

de Moran (Figura 11) y la identificación de una estadística importante de Moran I (0,1508; p<0,001), se encuentra una clara evidencia de autocorrelación espacial en la temperatura en Colombia, en el año 2015.

Contrastes de dependencia espacial local univariante

Estos contrastes, permiten identificar clústeres o asociaciones significativas de valores altos y bajos de una observación, con respecto a sus regiones vecinas. Por otra parte, determinan si el esquema de auto-correlación especial, es constante en todo el espacio geográfico objeto de estudio o si, por el contrario, la dependencia espacial es una tendencia grupal en la que existe la posibilidad de que se localicen conjuntos de observaciones, que no evidencien ningún esquema de asociación espacial.

Mapas LISA (indicadores de asociación espacial local)

Los indicadores de asociación espacial local, se pueden representar gráficamente como mapas que determinan localizaciones con valores significativos. Asimismo, permiten el cálculo del estadístico I de Moran de asociación espacial local, es decir, para cada observación se calcula un estadístico I. La sumatoria de los estadísticos Ii de todas las observaciones, darán como resultado el estadístico de asociación espacial global (donde los criterios de contigüidad definen la relación de vecindad entre unas observaciones y las posibles regiones vecinas, es decir, los esquemas de autocorrelación espacial local variarán en función del criterio de contigüidad, utilizándose en general el criterio de la reina, que contempla el análisis multidimensional más completo).

A continuación, se presenta el mapa de clústeres o asociaciones espaciales significativas locales para la variable temperatura promedio en los departamentos de Colombia. Ver figura 12.

Como se observa, existen esquemas de asociación espacial local para la temperatura con valores significativos altos (tono rojo), en los departamentos de La Guajira, Cesar y

Tabla 1. *Estadísticos LISA para la temperatura, en el periodo de 2015, por departamentos en Colombia.*

	CODDEPT	LISA_1		CODDEPT	LISA_I
1	Amazonas	0,2643222	18	Guaviare	0.0784926
2	Antioquía	0,0011932	19	Huila	-0,7730966
3	Arauca	-0,3582643	20	Magdalena	0,7750324
4	Atlantico	0,5306621	21	Meta	0,1046470
5	Bogota	1,3472577	22	Nariño	0,7446129
6	Bolivar	0,2072494	23	Norte de Santander	-0,3582643
7	Boyaca	0,0060941	24	Putumayo	-0,4015276
8	Caldas	0,6078940	25	Quindio	0,0000102
9	Caqueta	0,0645673	26	Risaralda	0,0003482
10	Casanare	-0,3024275	27	San Andres y Providencia	0,0000000
11	Cauca	0,0487942	28	Santander	-0,0174303
12	Cesar	0,5306621	29	Sucre	-0,8916202
13	Choco	-0,0020177	30	Tolima	0,0016519
14	Cordoba	-0,1516368	31	Valle del Cauca	0,0004158
15	Cundinamarca	1,1624590	32	Vaupes	0,3142609
16	Guajira	0,7973401	33	Vichada	0,0564554
17	Guania	0,2421272			

Fuente: *Elaboración propia a partir de GeoDa 1.0.1., 2016.*

Magdalena y las correspondientes localizaciones vecinas de cada uno; valores significativos bajos (tono azul), para el departamento de Cundinamarca; valores poco significativos –bajo alto– (tono azul claro) para el departamento de Sucre, es decir, los departamentos contiguos a Sucre presentan valores de temperatura promedio relativamente altos con respecto a ese departamento; y, valores poco significativos –alto bajo–(tono rosado) para el departamento del Huila, o sea, los departamentos contiguos a Huila presentan valores de temperatura promedio bajos con respecto a dicho departamento. Por su parte, San Andrés no tiene regiones vecinas con las que pueda asemejarse a algún patrón de asociación espacial; al respecto, para efectos de un mejor análisis especial, en el que se presente este tipo de situaciones y existan observaciones muy distanciadas o excluidas del resto de observaciones, como es el caso, se hace necesario definir un criterio de contigüidad que se base en el reconocimiento de distancias mínimas d, y no de la determinación de un límite o frontera común.

La sumatoria de los estadísticos de asociación local (tabla 1), dan como resultado el valor del estadístico de auto-correlación espacial global I de Moran, que es de 0,1508. La importancia de identificar los esquemas de auto-correlación espacial radica, principalmente como un criterio, en la selección del modelo econométrico, puesto que, si existe una situación bien definida de auto-correlación espacial, es necesario optar por un modelo econométrico espacial que recoja la información procedente de la auto-correlación (donde las fuentes de autocorrelación en un modelo se deben, principalmente, a la existencia de esquemas de tendencias bien definidos, variables incorrectamente especificadas, omisión de variable y retardos espaciales o temporales), y la replique como una variable más dentro del modelo.

Conclusiones

El AEDE, es una técnica descriptiva que permite la combinación de herramientas de la rigurosidad estadística, con métodos de análisis gráficos, lo cual ayuda a identificar y analizar la estructura de la distribución espacial en un contexto univariante o multivariante, al determinar cómo es la regularidad en los esquemas de asociación espacial, cuando las variables en cuestión no tengan un referente hipotético, que sugiera alguna idea del comportamiento de las variables.

El AEDE, debe constituir el primer eslabón en un análisis modelizador y decisor en el campo de la investigación ambiental, social y económica. En este artículo, se han presentado las principales técnicas del AEDE, que combinan el análisis estadístico con el análisis gráfico para hacer posible el estudio de las distribuciones espaciales y sus valores atípicos, esquemas de asociación espacial y agrupamientos espaciales.

Se concluye que, en el análisis de las series geográficas, se requiere de herramientas propias, que van más allá de las convencionales técnicas del AED o minería de datos y, por tanto, de un software específico. Estas herramientas, deben estar dirigidas al análisis de dos elementos fundamentales: tendencia espacial y puntos atípicos. Esto último, se entiende no solo como la determinación de valores significativamente altos o bajos de una variable, sino también como "concentración" de valores similares o disimilares en torno a una unidad geográfica (dependencia espacial).

El AEDE entonces, constituye la etapa previa al modelamiento econométrico espacial, porque, finalmente, en esta fase se determina si existe la necesidad de aplicar econometría espacial o si, por el contrario, los métodos incluidos dentro de la econometría convencional siguen siendo útiles.

Los mapas temáticos de cuantiles, son útiles para determinar la tendencia espacial; sin embargo, el riesgo de interpretación inadecuada de los datos y la obtención de resultados espurios, se hace mayor cuando no es posible clasificar los datos en rangos de cuantiles, debido a la gran diversidad en los valores de los datos.

Los mapas de desviación típica,

permiten una mejor identificación de la tendencia espacial, pues no clasifican los datos con respecto a los rangos de cuantiles, si no que establecen unas categorías más precisas, basadas en el grado de dispersión de los valores de los datos con respecto a la media.

En el contexto multivariante, el análisis del gráfico de coordenadas paralelas permite identificar la tendencia de una variable asociada con múltiples variables; en consecuencia, es posible determinar un valor medio esperado condicional en situaciones de estudio, en el que se presentan más de dos variables.

Referencias

ACEVEDO, I., & Velásquez, E. (2008). Algunos conceptos de la econometría espacial y el análisis exploratorio de datos espaciales. *Ecos de Economía*, (27), 9-34.

ANSELIN, L. (1995). Local Indicators of Spatial Association-LISA. *Geographical Analysis*, 27(2), 93-115.

BELLEHUMEUR, C., & Legendre, P. (1998). Multiscale Sources of Variation in Ecological Variables: Modeling Spatial Dispersion, Elaborating Sampling Designs. *Landscape Ecology*, (13), 15-25.

BIONDI, F., Myers, D., & Avery, C. C. (1994). Geostatistically Modeling Stem Size and Increment in an Old-Growth Forest. *Cananadian Journal of Forest Research*, 24, 1354-1368.

BRADSHAW, G. A., & Spies, T. A. (1992). Characterizing Canopy Gap Structure in Forests Using Wavelet Analysis. *The Journal of Ecology*, 80(2), 205-215.

CASETTI, E., & Poon, J. (1995). Econometric Models and Spatial Parametric Instability. Relevant Concepts and an Instability Index. En L. Anselin & R. Florax (Eds.), *New Directions in Spatial Econometrics* (pp. 301-321). Berlin, Germany: Springer.

CHASCO YRIGOYEN, C. (2003). *Métodos gráficos del análisis exploratorio de datos espaciales*. Madrid: Instituto L. R./Universidad Autónoma de Madrid.

CRESSIE, N. (1993). *Statistics for Spatial Data* (Revised edition). New York: Wiley.

CRESSIE, N. (2001). Fitting Variogram Models of Weighted Least Squares. *Journal of the International Associarion of Mathematical Geology*, 17, 563-86.

DALE, M. (1999). *Spatial Pattern Analysis in Plant Ecology*. Cambridge: Cambridge University Press.

DELFINER, P. (1979). *Basic Introduction to Geostatistics*. Paris: Ecole des Mines.

DUNCAN, R. P. (1991). Competition and the Coexistence of Species in a Mixed Podocarp stand. *Journal of Ecology*, 79(4), 1073-1084.

DYKES, J. (1998). Cartographic Visualization: Exploratory Spatial Data Analysis with Local Indicators of Spatial Association Using Tcl/Tk and CDV. *The Statistician*, 47(3), 485-497.

EPPERSON, B. K., & Li, T.-Q. (1996). Measurement of Genetic Structure within Populations Using Moran's Spatial Autocorrelation Statistics. *Proceedings of the National Academy of Sciences the USA*, (93), 10528-10532.

GRAHAM, D. J., & Glaister, S. (2003). Spatial Variation in Road Pedestrian Casualties: The Role of Urban Scale, Density and Land-use Mix. *Urban Studies*, 40(8) , 1591-1607.

HAINING, R. S. (2000). Providing Scientific Visualization for Spatial Data Analysis: Criteria and an Assessment of SAGE. *Journal of Geographical Systems*, 2, 121-140.

JOURNEL, A. G., & Huijbregs, Ch. (1978). *Mining Geostatistics*. New York: Academic Press.

LEDUC, A., Drapeau, P., Bergeron, Y., & Legende, P. (1992). Study of Spatial Components of Forest Cover Using Partial Mantel Tests and Path Analysis. *Journal Vegegetation Science*, 3(1), 69-78.

MORENO, R., & Vayá, E.. (2000). *Técnicas econométricas para el tratamiento de datos espaciales: La econometría espacial*. Barcelona: Edicions Universitat de Barcelona.

OKABE, A., Satoh, T., & Sugihara, K. A. (2009). A Kernel Density Estimation Method for Networks, its Computational Method and GIS-Based Tool. *Geographic Information Science*, 23(1), 7-32.

ORD, J. (1975). Estimation Methods for Models of Spatial Interaction. *Journal of the American Statistical Association*, 70, 120-126.

PHILLIPS, J. D. (1985). Measuring Complexity of Environmental Gradients. *Vegetatio*, 64(2-3), 95-102.

SAMPSON, R. (1987). Urban Black Violence: The Effect of Male Joblessness and Family Disruption. *American Journal of Sociology*, (93), 348-382.

SMOUSE P., Long, J. C., & Sokal, R. R. (1986). Multiple regression and correlation extension of the Mantel test ofmatrix correspondence. *Systmatic Zoology*, (35), 627-632.

UNWIN, A. (2000). Using Your Eyes-Making Statistics More Visible With Computers. *Computational Statistics & Data Analysis*, (32), 303-312.

VAYÁ, E., Moreno, R., & Suriñach, J. (2002). Economic Growth and Spatial Externalities. En L. Anselin, R. Florax, & S. Rey (Eds.), *Advances in Spatial Econometrics* (pp.145-156), Springer-Verlag: Heidelberg.

VELLEMAN, P. F. (1981). A*pplications, basics, and computing of exploratory data analysis*. Boston: Duxbury.

VER HOEF, J. M., Cressie, N., & Glenn-Lewin, D. (1993). Spatial models for spatial statistics: some unification. *Journal of Vegetation Science*, 4, 441-452. doi: 10.2307/3236071

VILALTA PERDOMO, C. J. (2005). Cómo enseñar autocorrelación espacial. *Economia, Sociedad y Territorio*, 18, 323-333.

WAGNER, H. H. (2003). Spatial covariance in plant communities: integrating ordination, geostatistics, and variance testing. *Ecology*, 84, 1045-1057.

WARRICK, A. W. & Myers, D. E. (1987). Moptimization of Sampling Locations for Variogram Calculations. *Water Resources Research*, (23), 496-500.

WARTENBERG, D. (1985). Multivariate spatial correlation: a method for Exploratory Geographical Analysis. *Geographical Analysis*, 17(4), 263-283.

WHITTLE, P. (1954). On Stationary Processes in the Plane. *Biometrika*, (41), 434-449.