Análisis de estadística espacial para el Índice de Desarrollo Social Cantonal de Costa Rica.

Melissa Cordero, Marco Otoya

Universidad de Costa Rica. 11 de diciembre, 2019

Resumen

El presente documento realiza un análisis de estadística espacial de áreas aplicado al Índice de Desarrollo Social Cantonal de Costa Rica, 2017; el índice permite ordenar los distritos según su nivel de desarrollo social. Para el análisis se utilizaron variables socio económicas estimando un modelo de regresión lineal y comparándolos con la estimación de un modelo CAR y SAR. Se encontró evidencia estadística a favor de la autocorrelación espacial para el IDS cantonal, un modelo de tipo SAR logra un mejor ajuste con respecto al modelo de regresión lineal y al modelo CAR.

1. Introducción

Este estudio realiza un análisis desde el enfoque de estadística espacial del Índice de Desarrollo Social Cantonal (IDS). El IDS cantonal aborda condiciones esenciales para el desarrollo social en dimensiones como educación, salud, participación ciudadana, económicas y seguridad según lo indica la Declaración Universal de los Derechos Humanos (MIDEPLAN, 2018).

En este contexto el índice es una herramienta del Estado para canalizar recursos hacia los distintos cantones que tienen un acceso limitado en diferentes dimensiones (económica, social, ambiental), por lo que analizar y conocer el componente espacial asociado contribuye a canalizar las acciones con el fin de mejorar las condiciones económicas y sociales de la población.

El índice es publicado por el Ministerio de Planificación Económica (MIDEPLAN) y se encuentra disponible para el año 2017 por cantón y distrito. De acuerdo con MIDEPLAN (2017), el desarrollo social se refiere a la calidad de vida de los seres humanos y su entorno; desde este punto de vista el concepto incluye una serie de consideraciones que intentan dar una visión más amplia y tienen que ver con nociones de pobreza, necesidades básicas, vulnerabilidad, sostenibilidad, solidaridad, igualdad, equidad, libertad, bienestar, exclusión social y seguridad, entre otros

El Índice incorpora cinco dimensiones: a-) Económica que se refiere a la participación en la actividad económica y gozar de condiciones adecuadas de inserción laboral que permitan

un ingreso suficiente para lograr un nivel de vida digno; b-) Participación social reflejada en los procesos cívicos nacionales y locales, para que se desarrolle en la población el sentido de pertenencia y de cohesión social y con ello el sentimiento de participación activa responsable, que implica el deber y el derecho de los ciudadanos a participar en los mismos; c-) Salud: orientada a gozar de una vida sana y saludable, lo que implica contar y tener acceso a redes formales servicios de salud, así como a una nutrición apropiada, que garanticen una adecuada calidad de vida de la población, d-) Educativa: relacionada con la disponibilidad y el adecuado acceso de la población a los servicios de educación y capacitación que favorezcan un adecuado desarrollo del capital humano, e-) Seguridad: analizada desde la condición básica para que las personas puedan desarrollar sus capacidades, vivir y desenvolverse en un entorno libre de situaciones de violencia y delito que amenazan su integridad física.

Al trabajar con datos de carácter espacial, no siempre se cumple el supuesto de independencia que asumen la teoría estadística. En el caso de los datos espaciales cuanto más próximas estén dos observaciones pareciera lógico suponer que van a estar relacionadas. A esta falta de independencia se denomina autocorrelación espacial (Reglero, 2018); el análisis busca determinar si existe autocorrelación espacial para el caso del IDS.

La presente investigación tiene como objetivo determinar la existencia de un componente espacial en el índice de desarrollo social cantonal (IDS) de Costa Rica, de acuerdo con variables de medición del bienestar social. En este sentido interesa conocer si ¿existe autocorrelación espacial para el IDS?, ¿qué variables se pueden incorporar para analizar el IDS? Tanto desde un modelo de regresión lineal simple como de un modelo que incorpore el componente espacial.

2. Datos y métodos

2.1 Datos

Dado que el IDS se construye a partir de sus dimensiones, las dimensiones del índice no se consideran para el análisis. De manera alternativa se utilizan variables de tipo socioeconómicas y disponibles para los cantones, entre ellas, el porcentaje de hogares con acceso a internet, la población por cantón y la severidad de pobreza, obtenidas del Instituto Nacional de Estadística y Censos, adicionalmente, se consideró el consumo se de energía eléctrica por cantón, cuyos datos fueron obtenidos de la Autoridad Reguladora de los Servicios Públicos (Aresep).

2.2 Procedimientos

La presente investigación consideró el siguiente análisis:

a. Análisis gráfico del índice de desarrollo social para el país.

b. Creación de la base de datos y unión con el archivo “shape” de Costa Rica para los cantones.
c. Determinación de criterios con el propósito de definir vecinos para los cantones de Costa Rica a nivel del IDS.
d. Selección de la una matriz de pesos para el modelado espacial.
e. Análisis de la correlación espacial mediante el test de la I de Moran.
f. Análisis de conglomerados por medio gráfico para el IDS.
g. Modelar el IDS mediante un modelo de regresión lineal simple.
h. Modelar el IDS incorporando el componente espacial mediante modelos SAR y CAR
i. Realizar una comparación entre modelos para seleccionar el que mejor se ajusta al fenómeno que se desea explicar.

2.2 Metodología a utilizar

Los datos geoespaciales, se refieren a los datos que obtenemos sobre lugares geográficos. El uso de este tipo de datos reconoce “el rol clave que conceptos espaciales como la distancia, la ubicación, proximidad, vecindario y región juegan en la sociedad humana”, lo que permite abordar los fenómenos desde una perspectiva multivariada y multidimensional (Urdinez & Cruz Labrín).

La inferencia a partir de datos espaciales supone que las observaciones en el espacio no pueden ser siempre consideradas como independientes. La inferencia en base a datos espaciales supone que las observaciones que están juntas una a la otra es similares, siguiendo patrones espaciales o estando autocorrelacionados.

La autocorrelación se calcula a partir del Test de Moran:

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

El índice anterior mide la autocorrelación espacial global donde w_{ij}es el elemento (i; j) de la matriz de pesos$\overline{W}$, $\bar{y}$ es la media de la variable “Y” y “n” el número total de áreas.

El índice de Moran también puede estimarse a nivel local:

$$I_i = (y_j - \bar{y}) \sum_{j=1}^{N} w_{ij}(y_i - \bar{y}) \qquad (Ecuación\ 2)$$

El I de Moran local permite obtener un valor para cada unidad espacial, analizando el grado de dependencia individual de cada unidad respecto a las demás. Lo anterior permite identificar clúster o zonas de mayor similitud.
Con el propósito de incorporar el componente espacial y modelar la autocorrelación espacial, se consideran el modelo autorregresivo simultáneo (SAR) y el modelo autorregresivo condicional (CAR). Estos modelos permiten incorporar la estructura de vecindad. La principal diferencia que presentan estos dos modelos es que los condicionales se utilizan cuando existe dependencia entre datos cercanos mientras que los modelos SAR tienen cuenta la dependencia de mayor alcance (Reglero, 2018).

Para modelar las interacciones espaciales que surgen de datos espaciales referenciados, se incorpora la dependencia espacial en la estructura de covarianza de modelos autorregresivos. A partir del análisis de regresión, si la ubicación de las regiones de una zona geográfica es conocida, es común suponer que tales observaciones en regiones cercanas entre sí pueden tener puntuaciones similares en las variables omitidas en la regresión, causando que el término de error este espacialmente autocorrelacionado. Dado esto, es posible incluir un proceso o componente espacial subyacente en el modelo y a partir de una estructura de vecinos utilizar un modelo autorregresivo (SAR). El modelo SAR se puede especificar de la siguiente manera:

$$y(S_i) = \emptyset \sum_{j=1}^{n} W_{ij}\, y(S_j) + \epsilon \qquad (Ecuación\ 3)$$

$$y \sim \mathcal{N}(0, \sigma^2((1-\emptyset W)^{-1})((1-\emptyset W)^{-1})^t) \qquad (Ecuación\ 4)$$

Un segundo modelo que permite la incorporación de elementos espaciales en la regresión es el modelo autorregresivo condicional (CAR), que se específica de la siguiente forma.

$$y(S_i)|y - S_i \sim \mathcal{N}\left(\emptyset \sum_{j=1}^{n} W_{ij}\, y(S_j), \sigma^2\right) \qquad (Ecuación\ 5)$$

$$y \sim \mathcal{N}\left(0, \sigma^2((1-\emptyset W)^{-1})\right) \qquad (Ecuación\ 6)$$

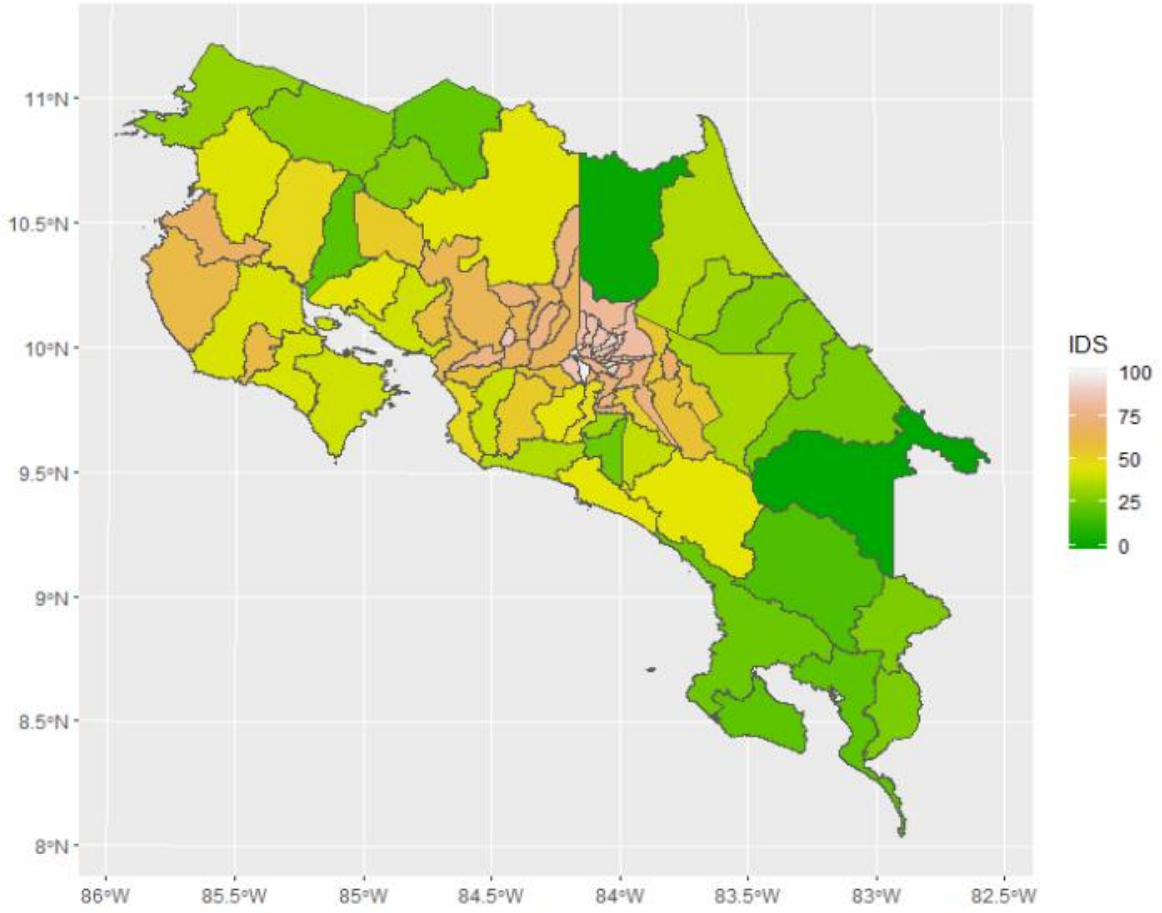
3. Resultados

3.1 Análisis descriptivo

El índice de desarrollo social cantonal esta disponible para los 82 cantones del país y asume valores que van de 0 a 100, su valor depende de la puntuación que asume cada una de las

dimensiones que lo componen. En la Figura 1 se muestra el IDS por cantón, como se muestra el color verde oscuro representa los cantones con un menor IDS particularmente se observan cantones como Sarapiquí y los cantones de la zona sur con valores bajos. Los cantones del centro del país y algunos como Liberia y Nicoya presentan IDS elevados, particularmente destaca en blanco el cantón de Escazú cuyo valor del IDS es 100.

Figura 1. Índice de Desarrollo Social por cantones.



Fuente: Elaboración propia con base en datos de MIDEPLAN.

3.2 Análisis de vecinos

Se requiere determinar el set de vecindarios para cada observación, es decir, identificar los polígonos que comparten fronteras entre ellos. A partir de lo anterior se asignarán pesos a cada relación vecina, de tal forma que permita definir la fuerza de esta relación en base a cercanía. En las matrices de peso, los vecinos son definidos por un método binario (0,1) en cada fila indicando si existe o no relación.

Se consideran tres criterios diferentes para calcular los vecindarios. El primer criterio es el de la "torre" (Rook) que considera como vecinos a cualquier par de celdas que compartan alguna arista (borde). Un segundo criterio es el método de la "Reina" (Queen) que considera como vecinos a cualquier par de celdas que compartan alguna arista o un punto. La Figura q 2 y 3 muestra la determinación e vecinos mediante el método de torre y reina respectivamente.

Figura 2. Criterio torre

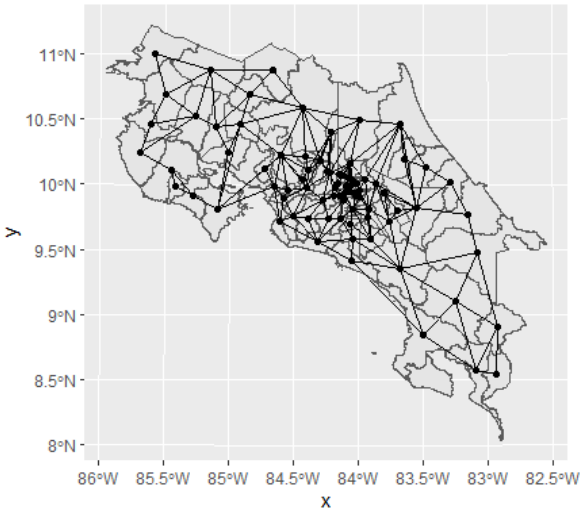
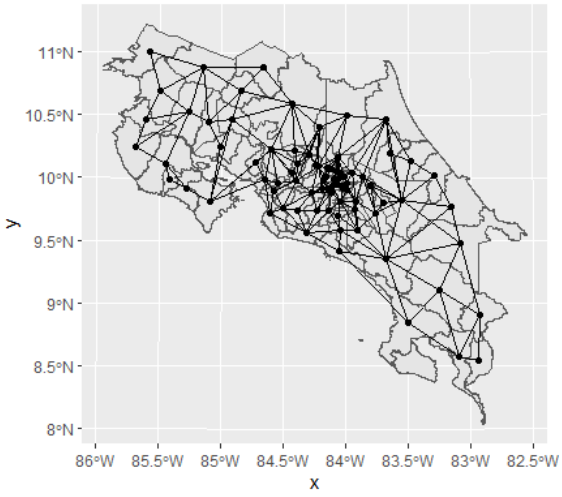


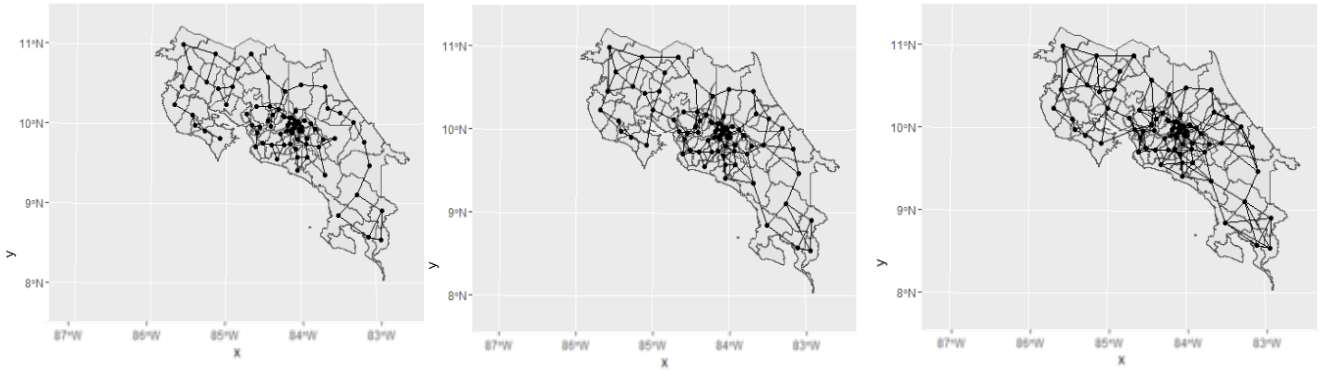
Figura 3. Criterio Reina



Fuente: Elaboración propia

Un tercer criterio es el de vecinos cercanos (K-nearest), en este caso los vecindarios se generan en base a la distancia entre vecinos, donde "k" se refiere al número de vecinos de una determinada locación, calculada como la distancia entre los puntos centrales de los polígonos. La figura 4 muestra los resultados con 2, 3 y 4 vecinos respectivamente.

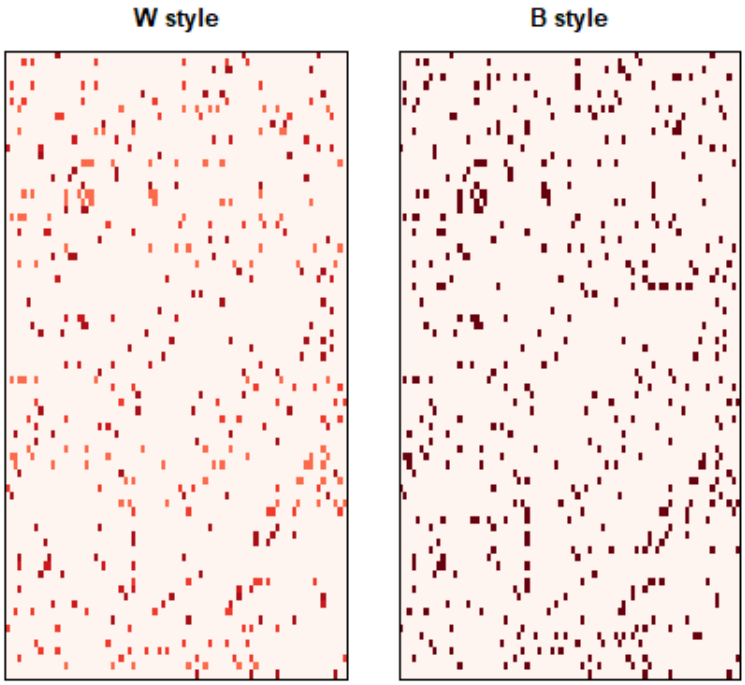
Figura 4. Criterio 2, 3 y 4 vecinos



Fuente: Elaboración propia.

Los resultados fueron similares entre distancias, torre y reina, por lo que se decidió emplear para determinar la matriz de pesos el criterio de la reina. Sin embargo, es claro que en el método de k-vecinos conforme se aumenta el número de vecinos se obtienen mayores relaciones.

Figura 5. Matriz de pesos por estilo.



Fuente: Elaboración propia.

Como se muestra en la Figura 5, los resultados son bastante similares entre todos los dos estilos, pero sobre todo se observa, por lo que se considera adecuado emplear el estilo W con la ventaja que logra una adecuada estandarización para cada cantón al restringir que la suma de todos los valores de 1. Adicionalmente se revisó que no existieran cantones con cero relaciones. No se obtuvieron resultados para el estilo inverso, por lo que no se obtuvieron ponderaciones o fueron 0 en este caso.

3.2 Análisis de correlación espacial

El estudio de autocorrelación espacial global busca detectar la presencia de tendencias o estructuras espaciales generales en la distribución de la variable sobre el espacio geográfico completo. Y saber si la variable se distribuye de forma independiente o, si por el contrario, existe algún tipo de asociación entre regiones vecinas (Reglero, 2018). Se utilizó el test de la I de Moran para confirmar la correlación espacial significativa.

Los resultados del test de I Moran I (Cuadro 1) muestran que existe una relación de correlación positiva, versus una expectativa de una leve relación negativa. El test resulta estadísticamente significativo al tener un valor-p menor a una significancia del 0.05. Por ende, IDS presenta altos grados de autocorrelación espacial a nivel de cantones.

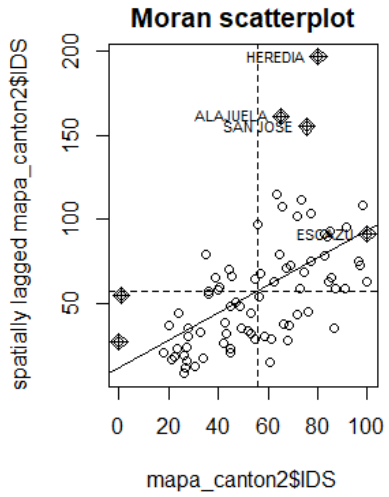
Cuadro 1. Resultados test I Moran.

Moran I test under randomisation			
data: mapa_canton2$IDS			
weights: cantones_lw_W			
Moran I statistic standard deviate = 9.9961	p-value < 2.2e-16		
alternative hypothesis: greater			
sample estimates:			
	Moran I statistic	**Expectation**	**Variance**
	0.671710982	**-0.0125**	**0.00468508**

Fuente: Elaboración propia.

Adicionalmente, la figura 6 visualiza el tipo y fuerza de la correlación espacial en la distribución de nuestros datos. La curva del gráfico indica el valor del Moran's I, es decir, la medida global de autocorrelación espacial de los datos, la cual es positiva. El eje horizontal se muestra el IDS a por cantón y el eje vertical muestra estos mismos datos, pero rezagados espacialmente. Los cuatro cuadrantes del gráfico describen el valor de la observación en relación con sus vecinos: alta-alta, baja-baja (autocorrelación espacial positiva), baja-alta o alta-baja (correlación espacial negativa). También es posible observar valores outliers en esta relación como Heredia, Alajuela y San José.

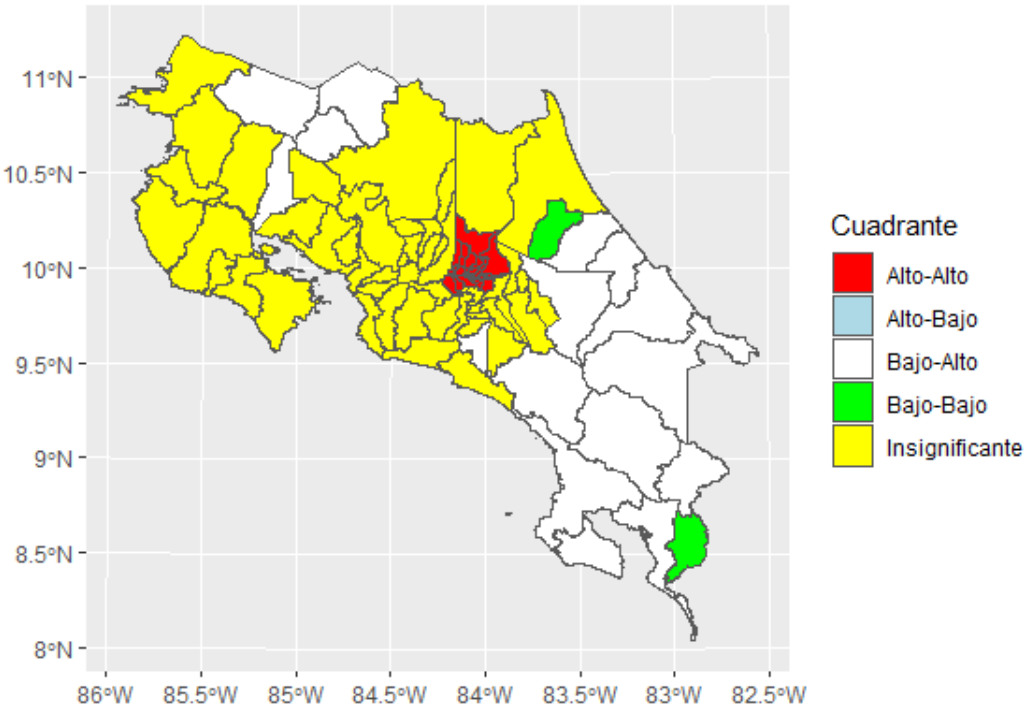
Figura 6. Resultados test I Moran.



Fuente: Elaboración propia.

La figura 7, muestra los patrones de clusterización para distintos valores del IDS, permite observar la variación en la autocorrelación a lo largo del espacio, pero no es posible identificar si los patrones geográficos de autocorrelación son clusters con valores altos o bajos, lo que nos permitirá analizar el tipo de autocorrelación espacial que existe y su nivel de significancia. Este mapa nos muestra si es que existen clusters, es decir, regiones en donde en su núcleo existe autocorrelación espacial positiva, por ende, regiones clusterizadas más que lugares individuales.

Figura 7. Patrones Geográficos de Clusterización para distintos valores de IDS

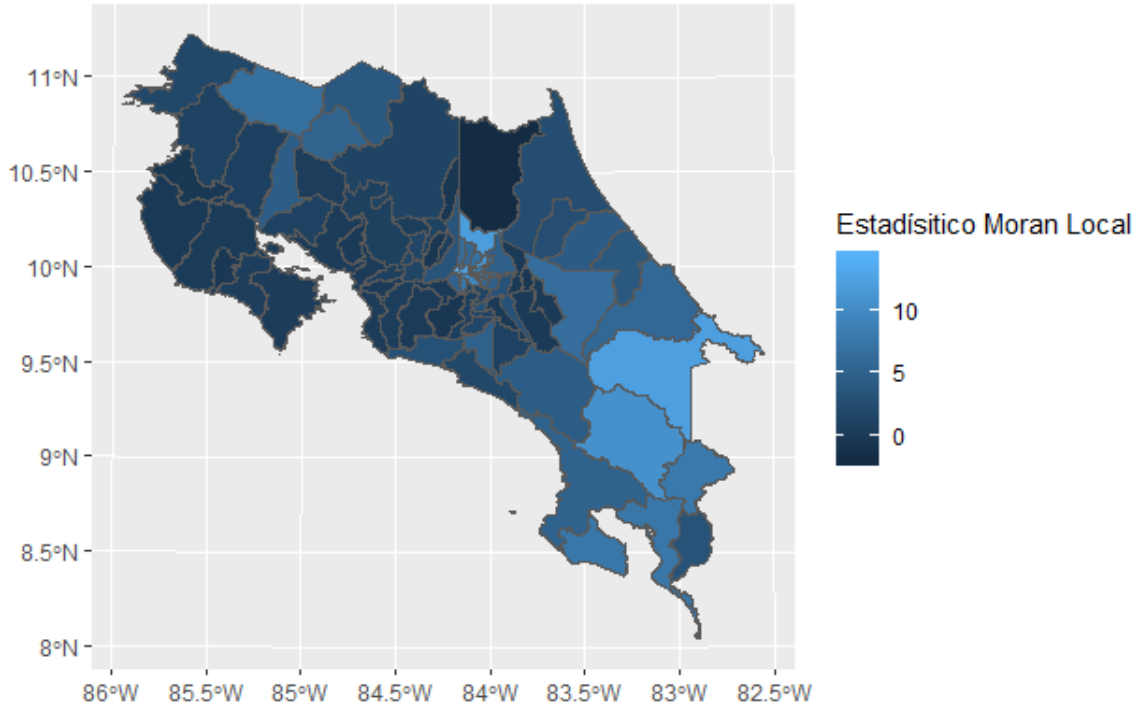


Fuente: Elaboración propia.

La zona sur tiene una correlación espacial baja-alta, es decir, su valor esta influenciada por la proximidad entre valores similares. Los cantones del centro del país tienen una correlación espacial alta, mientras que en el caso de los cantones en amarillo se muestra que dicha correlación es insignificante.

Finalmente, a nivel local se muestra en la figura 8 los valores del estadístico de Moran.

Figura 8. Estadístico Moran local



Fuente: Elaboración propia.

3.3 Análisis de regresión lineal

El modelo de regresión lineal para el IDS se realiza considerando las variables de consumo eléctrico, población, severidad de pobreza, porcentaje de hogares con acceso internet. Los coeficientes para las dos primeras variables resultaron no significativos. El grado de ajuste del modelo mediante el R cuadrado ajustado es de 73%.

Cuadro 2. IDS modelo de regresión lineal.

lm(formula = IDS ~ C_Elec + Severidad_Pobreza + Hog_Acc_Intern + Poblacion, data = mapa_canton2)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.76E+01	6.99E+00	5.376	8.08E-07	***
C_Elec	-9.14E-09	9.97E-09	-0.917	0.362	
Severidad_Pobreza	-2.41E+00	7.53E-01	-3.2	0.002	**
Hog_Acc_Intern	9.92E-01	1.40E-01	7.095	5.82E-10	***
Poblacion	6.54E-06	2.64E-05	0.248	0.805	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

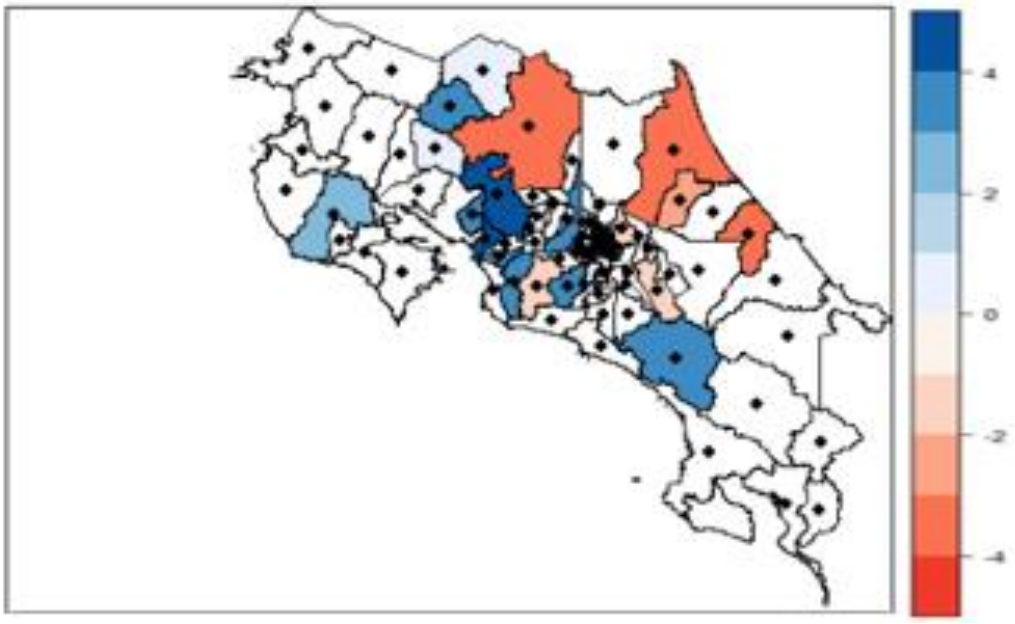
Multiple R-squared: 0.7463, Adjusted R-squared: 0.733

F-statistic: 55.9 on 4 and 76 DF, p-value: < 2.2e-16

Fuente: Elaboración propia.

El test de Moran global para los residuos de la regresión muestra una correlación positiva. La figura 9 muestra los residuos espacialmente, en las áreas color blanco, al parecer casi no hay errores de medición. En esos residuos para el modelo estimado no hay un componente espacial. Hay residuos grandes en ciertas regiones lo que indica la posibilidad de que ciertos elementos espaciales no se estén capturando.

Figura 9. Análisis espacial de los residuos.



Fuente: Elaboración propia.

3.4 Análisis de regresión espacial mediante modelos SAR y CAR

El cuadro 3 muestra los resultados para la regresión del IDS mediante el modelo SAR, como se observa los coeficientes de las variables resultan estadísticamente significativos y con los signos esperados. Posteriormente se ajustó un modelo ponderando por la población, sin embargo, los resultados sugieren que el patrón que detectamos antes puede deberse a un patrón espacial, y no el efecto del tamaño de la población en las áreas.

Cuadro 3. IDS modelo de regresión SAR.

Call: spautolm(formula = IDS ~ Severidad_Pobreza + Hog_Acc_Intern,

Residuals:

Min	1Q	Median	3Q	Max
-34.80744	-6.98016	0.55318	6.84747	27.06487

Coefficients:

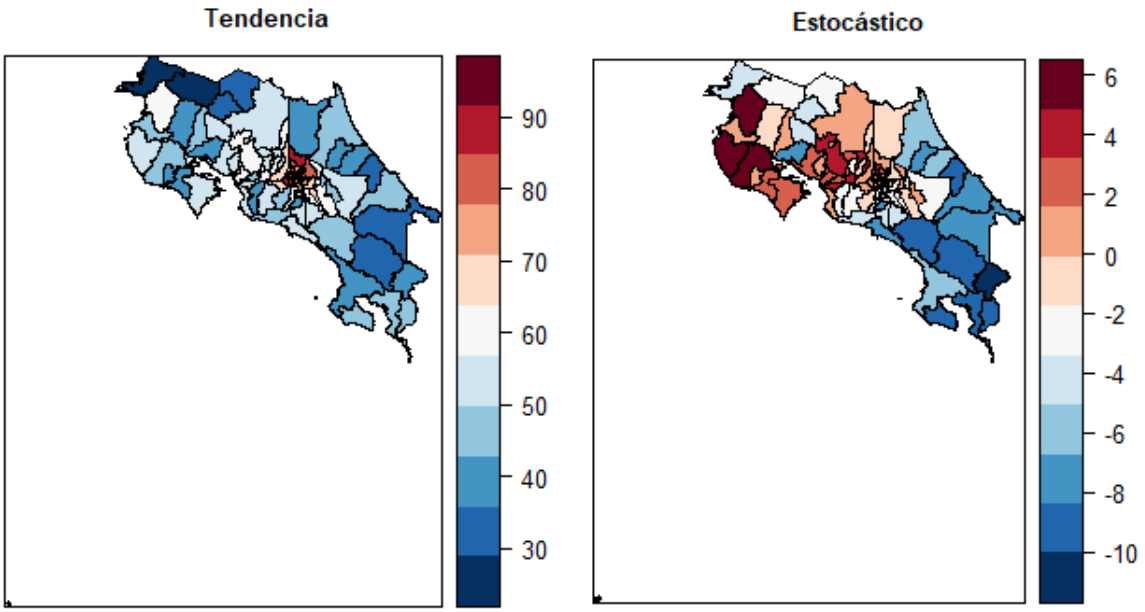
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	36.13395	8.11759	4.4513	8.54E-06
Severidad_Pobreza	-1.49404	0.83834	-1.7821	0.07472
Hog_Acc_Intern	0.9193	0.16258	5.6546	1.56E-08

AIC: 636.52

Fuente: Elaboración propia.

La figura 10 muestra los residuos para esta regresión, la grafica es diferente al modelo de regresión lineal, los valores de los residuos cambian en este caso algunos de ellos mayores con respecto al modelo de regresión lineal.

Figura 10. Análisis espacial de los residuos modelo SAR.



Fuente: Elaboración propia.

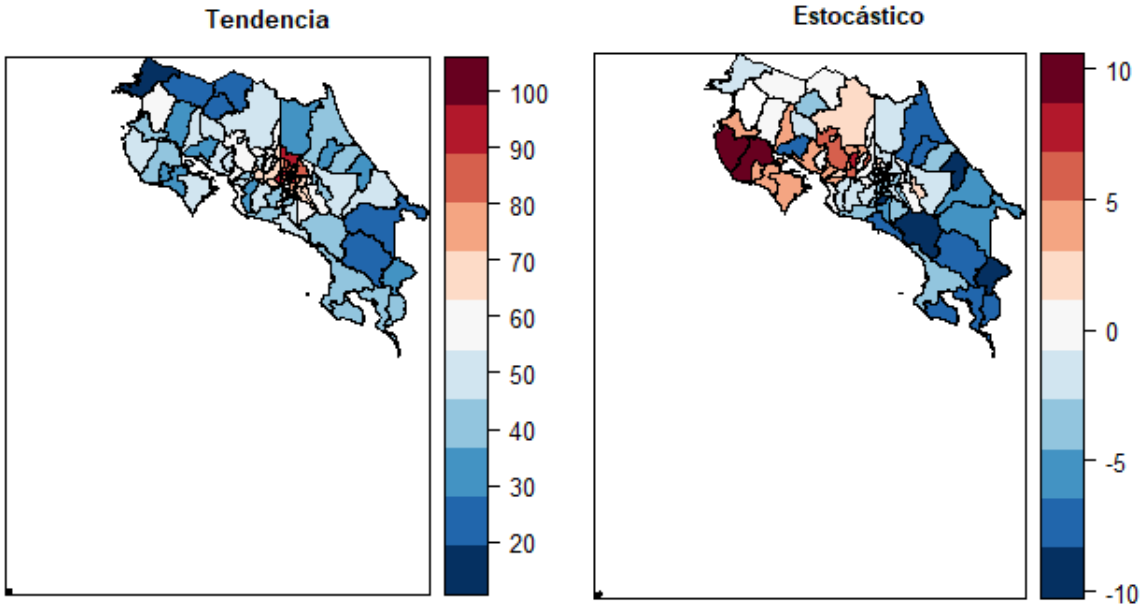
Finalmente, el cuadro 4 muestra los resultados para el modelo CAR, los resultados son similares al modelo SAR. En la Figura 11 se muestra el análisis de los residuos y no se observa una diferencia entre ambos modelos.

Cuadro 4. IDS modelo de regresión SAR.

Call: spautolm(formula = IDS ~ Severidad_Pobreza + Hog_Acc_Intern,					
Residuals:					
	Min	1Q	Median	3Q	Max
	-30.4152	-7.7893	1.3981	6.8285	28.5222
Coefficients:					
		Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)		33.68331	7.65734	4.3988	1.09E-05
Severidad_Pobreza		-2.0843	0.81144	-2.5686	0.01021
Hog_Acc_Intern		1.0669	0.15311	6.9683	3.21E-12
Lambda: 0.64484 LR test value: 6.5705 p-value: 0.010368					
Numerical Hessian standard error of lambda: 0.24265					
AIC: 638.99					

Fuente: Elaboración propia.

Figura 11. Análisis espacial de los residuos modelo CAR.



Fuente: Elaboración propia.

Dado que se genera evidencia de un componente espacial para el IDS se considera que un modelo espacial explica de mejor manera el IDS, por tanto, se compararon los tres modelos mediante el AIC, resultando el modelo SAR con el mejor indicador.

4. Conclusiones

Se determina que existe un componente espacial para el IDS a nivel cantonal, que internamente se explica a partir de la configuración económica y social de las regiones.

Los resultados del test de I Moran a nivel global, mostraron que existe una relación de correlación positiva. El test resulta estadísticamente significativo al tener un valor-p menor a una significancia del 0.05. Por ende, IDS presenta altos grados de autocorrelación espacial a nivel de cantones.

Existe evidencia de patrones de clusterización para distintos valores del IDS, permite observar la variación en la autocorrelación a lo largo del espacio. La zona sur tiene una correlación espacial baja-alta, es decir, su valor está influenciado por la proximidad entre valores similares. Los cantones del centro del país tienen una correlación espacial alta, mientras que en el caso de los cantones en amarillo se muestra que dicha correlación es insignificante.

El modelo de regresión lineal para el IDS se estimó considerando las variables de consumo eléctrico, población, severidad de pobreza, porcentaje de hogares con acceso internet. Los coeficientes para las dos primeras variables resultaron no significativos. El grado de ajuste del modelo mediante el R cuadrado ajustado es de 73%.

Dado la evidencia de autocorrelación espacial se estimaron modelos SAR y CAR, y se compararon mediante el valor del AIC. El modelo SAR

Un modelo estadístico para análisis de áreas SAR, logra explicar mejor IDS en comparación con un modelo que no toma en cuenta el componente espacial.

5. Enlace a repositorio

Para ver las bases de datos utilizadas y el código utilizado, por favor referirse a este enlace: https://github.com/posgrado-de-estadistica/proyecto4-melissa-y-marco/upload/master

6. Bibliografía

MIDEPLAN. (2018). *Índice de desarrollo social 2017.* San José, Costa Rica.

Reglero, C. M. (2018). Análisis de datos espaciales. Tesis para optar por el grado de Magister. Universidad de A Coruña. Coruña, España.

Urdinez, F., & Cruz Labrín, A. (s.f.). AnalizaR Datos Políticos. Santiago, Chile.