EVALUACIÓN DE MODELOS DE INTERPOLACIÓN PARA CÁLCULO DE AISLAMIENTO EN LA REGIÓN DE O'HIGGINS.

Matías Andrés Poch Clavero - Ingeniero Civil en Geografía.

icg.poch@yahoo.cl

RESUMEN.

El estudio "Identificación de Localidades en Condición de Aislamiento", realizado por la Subsecretaría de Desarrollo Regional (SUBDERE), se desarrolló un índice que permitiese identificar cuáles y cuántas localidades se encuentran en condición de aislamiento. El cálculo de este índice, estaba compuesto por la medición de tiempo de traslados de las localidades a una serie de servicios. El procedimiento para obtener los tiempos era ejecutar algoritmos de rutas mínimas en una red, llamada red de interconexión.

Para cada localidad (representada por un punto en el espacio) existe un índice de aislamiento, sin embargo, esta es una representación discreta en el espacio. Para este trabajo se plantea realizar diferentes métodos de interpolación, para un set de puntos (localidades de la región de O'Higgins) en que se ha medido el grado de aislamiento. Se plantea en generar un raster, que pueda representar de la mejor manera, el fenómeno medido de manera puntual, de manera que se puedan clasificar píxeles y así determinar "zonas de condiciones similares" o "territorios aislados".

Pero antes de realizar reclasificaciones para determinar zonas con índices similares, se deben utilizar medidas de ajustes de índole estadístico, para ver la confiabilidad de los métodos de interpolación. Se probarán 5 métodos de interpolación y rasterización, y se contrastarán los resultados con cuatro medidas de ajuste.

La región de O'Higgins cuenta con 2850 localidades con índice de aislamiento. Se elegirán al azar 339 puntos de control, quedando 2511 puntos para realizar las interpolaciones raster. Luego, se remuestrean los valores de los raster con los 339 puntos de control, se determina el mejor modelo de interpolación y se ejecuta la iterpolación con el mejor nivel de ajuste para las 2850 localidades.

PALABRAS CLAVES: interpolación, raster, aislamiento, remuestreo, medidas de ajuste.

1. INTRODUCCION.

Los modelos de datos raster, tiene como característica llevar a cabo una representación "discreta" del mundo real, utilizando una "grilla" de tamaño regular denominada "pixel". Cada pixel almacena un valor numérico que representa el valor de un determinado fenómeno del mundo real. Dentro de las características más generales del modelo de datos raster se encuentran:

- Regulares: Todos los píxeles son regulares, es decir todos iguales, estableciendo una codificación discreta de las coordenadas. El píxel es la mínima unidad, por lo que el tamaño de píxel dará la precisión con la que se pueden definir elementos geográficos. (resolución espacial)
- Localizables: La grilla cuenta con fila y columna. Todo píxel puede ser identificado por su posición (i,j) en la matriz de un raster, a partir de su número de fila y columna.
- Contiene un valor numérico que puede ser el identificador de la celda, valor temático,

Existen muchos fenómenos que son medidos mediante puntos discretos en el espacio y necesitan ser

representados de forma continua; por ejemplo, cantidades de contaminantes en agua o suelo. Para representar estos puntos discretos se utilizan métodos matemáticos conocidos como ***interpolación***.

Interpolar es el método matemático que permite hallar un valor o dato que no se ha obtenido directamente, en un intervalo donde solo se conocen los valores extremos. Sin embargo, no todos los métodos de interpolación sirven para hallar un dato o valor, es por eso que en este trabajo se plantea una serie de medidas de ajuste, en que, estadísticamente se demuestra qué tipo de interpolación es mejor para la explicación de un determinado fenómeno.

2. OBJETIVOS.

2.1. General.

Determinar que modelo de interpolación es adecuado para representar el fenómeno de aislamiento de las localidades en la región de O'Higgins.

2.2. Específicos.

- Revisión de los métodos de interpolación provistos por gvSIG 1.11.
- Generar 5 modelos raster, a partir de los métodos de interpolación.
- Calcular medidas de ajuste, mediante puntos de control.

3. METOLOGÍA.

3.1. Métodos de interpolación.

El Software gvSIG a través de su gestor de geoprocesos SEXTANTE, posee algoritmos para crear capas raster a partir de puntos. Los métodos de interpolación analizados en este trabajo son:

- Decrecimiento lineal:Este algoritmo crea una nueva capa raster continua a partir de una serie de valores puntuales, mediante el método de interpolación por decremento lineal.
- Densidad: Este algoritmo calcula el valor de densidad de puntos para cada celda de la capa raster de salida. La densidad representa la intensidad (propiedad de primer orden) del patrón de puntos, que es una realización concreta del proceso de puntos que lo ha generado. El cálculo de densidades utilizando un área de influencia fija de dos únicos puntos. Donde ambas áreas intersecan, la densidad es, lógicamente, mayor. Las celdas que no están en el área de influencia de ningún punto tienen un valor de densidad nulo. En el caso por ejemplo de las observaciones de una especie, la capa de densidad nos proporciona una medida de la probabilidad de encontrar esa especie en cada celda. El cálculo de estas probabilidades es la base para el desarrollo de modelos predictivos más complejos.
- Densidad (Kernel): Este algoritmo calcula el valor de densidad de puntos para cada celda de la capa raster de salida a partir de la función núcleo (kernel).
- Distancia inversa: Este algoritmo crea una nueva capa raster continua a partir de una serie de valores puntuales, mediante el método de interpolación por Distancia Inversa. El valor en una coordenada dada se calcula mediante una media ponderada de los puntos de influencia seleccionados (bien sea la selección por distancia o por número de éstos). Al ser un método basado en ponderación por distancia, sólo se tiene en cuenta el alejamiento, pero no la posición. Es decir, un punto situado a una distancia (x) hacia el norte tiene la misma influencia que uno situado a esa misma distancia (x) pero hacia el oeste.
- Vecindad: Este algoritmo crea una capa raster interpolando a partir de una capa de puntos mediante el método de vecino más próximo. El valor asignado a cada celda es el correspondiente al punto más cercano.

3.2. Salidas raster.

Todas las salidas raster de las interpolaciones tendrán las siguientes características:

- Se realizará la extensión del raster a partir de la extensión de la capa con la totalidad de las localidades, que se encuentran en Datum WGS 84 H 19S.
- El tamaño de salida de celda es de 500 mts.
- Se recortará el raster con el polígono de la región de O'Higgins. Utilizando gestor de geoprocesos SEXTANTE El polígono resultante, en este caso el área resultante corresponde a la región de O'Higgins.
- Se ajustarán los límites de valores a la escala real del índice de aislamiento que va desde -1 a 2. Donde los valores menores a 0, según el equipo investigador, son localidades en condición de aislamiento.
- Se ocupará la paleta de colores stern-special (7 clases).
    - Blanco [2.0]: Máximo índice, indica que existe nulo aislamiento.
    - Verde [1.22, 2[: Buenos niveles de integración.
    - Lila [0.51, 1.22[: Integración moderada.
    - Púrpura [-0.24, 0.51[: Zona de transición de integración leve a aislamiento moderado.
    - Violeta [-0.25, -0.24[: Aislamiento moderado.
    - Rojo [-0.83 , -0.25[: Aislamiento Grave.
    - Negro [-1, -0.83[: Aislamiento crítico o zona fuera de interpolación.

3.3. Puntos de control (Tamaño de muestra).

Para validar los datos observados con los resultados de interpolación, se deberá extraer una selección de puntos. Por lo que debe obtener una muestra que sea representativa, el número de puntos de control (tamaño de muestra), debe estar con márgenes de error que se deseen obtener.

Una muestra, cualquiera que sea su magnitud, debe ser representativa del universo de datos. Se debe tener en cuenta que los valores límites de los datos determinan el numero de datos a muestrear.

El tamaño de la muestra depende de las siguientes variables:

- El *nivel de confianza* o riesgo que se acepta de error para presentar resultados: lo que se desea es que en otras muestras semejantes los resultados sean los mismos o muy parecidos. También podemos denominarlo *grado o nivel de seguridad*. El nivel de confianza habitual es de 0,05 ($\alpha = 0,05$). El nivel de confianza está asociado a una determinada probabilidad de ocurrencia, en una distribución normal.
- La varianza estimada en los datos. A mayor diversidad esperada, o al menos posible, en los posibles valores hará falta un mayor número de observaciones en la muestra.
- El margen de error que estamos dispuestos a aceptar.

Como se conoce el número de localidades (observaciones) de la región de O'Higgins se calcula de la siguiente manera Fórmula 1:

$$n = \frac{N}{1 + \frac{(e^2(N-1))}{(z^{2pq})}}$$

Fórmula 1

Donde:
n: Tamaño de la muestra que se desea conocer.
N: Tamaño de la muestra.
e : Error muestral.
z : Valor correspondiente al nivel de confianza. Un nivel de confianza del 95% (también lo expresamos así: α = 0,05) corresponde a z = 1,96 *sigmas* o *errores típicos;* z = 2 *(dos sigmas)* corresponde a un 95,5% (aproximadamente, α = 0,045). Con z = 2.57 el nivel de confianza sube al 99% (nos equivocaríamos una vez de cada 100).
pq: Varianza de los datos. El suponer que p=q quiere decir que para escoger la muestra nos ponemos en la hipótesis de que en los datos existe la máxima diversidad posible y no se corre el riesgos de quedar cortos en el número de datos a muestrear. Este valor de pq (= 0.25) es válido (válido para calcular el tamaño de la muestra).

3.4. Medidas de Ajuste.

De la aplicación de cada uno de los métodos se evalúan sus medidas de ajuste, lo que permite determinar que método es el más idóneo. Las medidas de ajuste se basan en los datos obtenidos de la comparación de los observados y los modelos simulados para determinar la conducta del modelo. Los dos criterios sobre los cuales las medidas se juzgan aquí son: el grado para lo cual la sensibilidad atraviesa un rango de magnitudes de error, y la habilidad para comparar el modelo de ejecución por sobre resultados diferentes.

El primero de estos criterios indica al analista, en un sentido cualitativo, cuán bien un modelo se desempeña y a qué grado este desempeño difiere entre los modelos separados. El segundo criterio permite que los modelos sean comparados en diferentes situaciones o parámetros.

Para este caso ocuparemos cuatro medidas de calidad de ajuste. Tal evaluación se realiza empleando la comparación entre lo estimado y lo observado que adopta la siguiente notación:

[Xij°] = Valor estimado en coordenadas *i,j*.

[Xij] = Valor observado en coordenadas *i,j*.

$\overline{X}$= Valor medio observado.

n = Es el número de puntos de control.

Esta notación, es la que se ocupará para la formulación de las siguientes medidas de ajuste:

Coeficiente de Determinación (R^2)

Esta es uno de las medidas de evaluación más usadas pero ha sido considerada insensible en relación a los parámetros diseñados *Wilson et al (1969) and Black (1973)*. Esta cantidad indica que proporción de la variación total en la respuesta X_{ij} se explica con el modelo ajustado Fórmula 2: y Fórmula 3:

$$R^2 = \frac{Var\ explicada}{Var\ Tot\ X_{ij}}$$

Fórmula 2:

Se calcula $$R^2 = 1 - \frac{\sum_{ij} (X_{ij} - X^{\circ}_{ij})^2}{\sum_{ij} (X_{ij} - \bar{X})^2}$$

Fórmula 3:

Desviación estándar de residuos.

Openshaw y Connolly (1977) hicieron uso extensivo de esta medida en sus evaluaciones de funciones alternativas de detención. Principalmente mide la dispersión de datos Fórmula 4.

$$A=\sqrt{\frac{\sum_{ij}(X_{ij}-X_{ij})^2}{n^2-1}}$$

Fórmula 4

Error Absoluto Medio.

Es la medida más simple y más probada, donde compara como diferencia absoluta el valor real con el estimado, según la cantidad de zonas (*n*) que represente y se define como Fórmula 5 .

$$MABSERR=\frac{\sum_{ij}(|X_{ij}-X_{ij}^{\circ}|)}{n^2}$$

Fórmula 5

Valor absoluto medio.

Se define como Fórmula 6 .

$$MABSERR^{\circ}=\frac{1}{n^2}\sum_{ij}(|\frac{X_{ij}-X_{ij}^{\circ}}{\bar{X}}|)$$

Fórmula 6

Este es el radio del error absoluto medio en relación a la magnitud de intercambio medio.

La elección de estas medidas de ajuste por sobre otras se basa en que son igualmente sensibles a los cambios en el error, y tienen la propiedad de ser comparables entre diferentes matrices.

4. DESARROLLO.

4.1. Universo de datos para interpolar y muestrear.

La región de O'Higgins cuenta con 2850 localidades con índice de aislamiento.

Para este caso:

N = 2850 localidades de la región.

Nivel de confianza 95% z=1.96.

Error e =5%.

Y suponiendo p=q= 0.5

La muestra necesaria será de 339 localidades (puntos de control; verdes), por lo tanto se interpolarán 2511 localidades (puntos azules) para la región de O'Higgins Ilustración 1

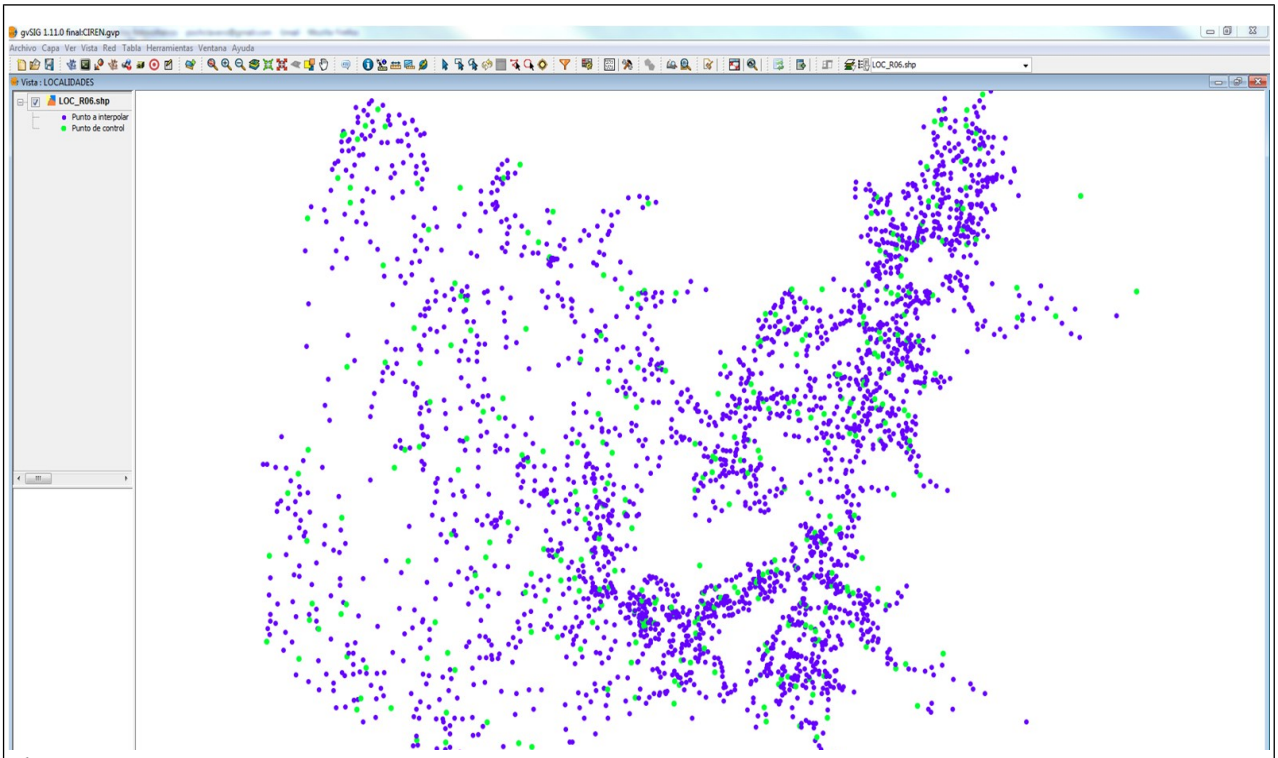

Ilustración 1

4.2. Interpolaciones.

Decrecimiento lineal. Ilustración 2

Parámetros:

- Radio de búsqueda: 5000 mts.
- Exponente: 2.

Se puede observar que este método a medida que se acerca a los límites espaciales de la interpolación, los valores son más homogéneos pareciendo grandes píxeles. Esto se debe a los radios de búsqueda en los extremos, dado que por la distancia menor a los 5 km esos píxeles tomarán el valor del punto interpolado.

La zona de mayor integración, corresponde a los espacios que se encuentran próximos a caminos. En cambio las zonas más violetas y rojas se ven distribuidas, preferentemente en la zona del secano costero.

Densidad.Ilustración 3

Parámetros:

- Radio de búsqueda: 5000 mts.

Debido a que el radio de búsqueda es un parámetro importante y que debe seleccionarse de acuerdo a la distribución de puntos, ya que condiciona los resultados. Lo que queda en manifiesto ya que con los parámetros introducidos para interpolar, no se logra discriminar, donde parecer ser inexistente el fenómeno de transición de ser integrado a aislados. Donde se sobreestima la representación de la integración.

Densidad (Kernel) Ilustración 4.

Parámetros.

- Radio de búsqueda: 5000 mts.

Arroja similares al modelo anterior, se ve mas suavizado debido a que a medida que se aleja del punto interpolado disminuye su influencia (y no como la misma en todos los puntos del radio de búsqueda).

Distancia inversa. Ilustración 5

- Parámetros:
- Radio de búsqueda: 5000 mts.

Exponente: 2.
Arroja resultados similares al decrecimiento lineal. Se observa, que los valores que adoptan los píxeles pasan de manera más "suave" de las zonas de mayor integración a las de aislamiento.

Vecindad. Ilustración 6

Por la naturaleza de este método (basado en el vecino mas cercano), se aprecia que la distribución del índice parece ser un mosaico proporcionalmente distribuido. El problema que tiene es que adopta zonas como homogéneas, a medida que el punto vecino más cercano este mas lejos estas zonas tienden a aumentar su extensión.

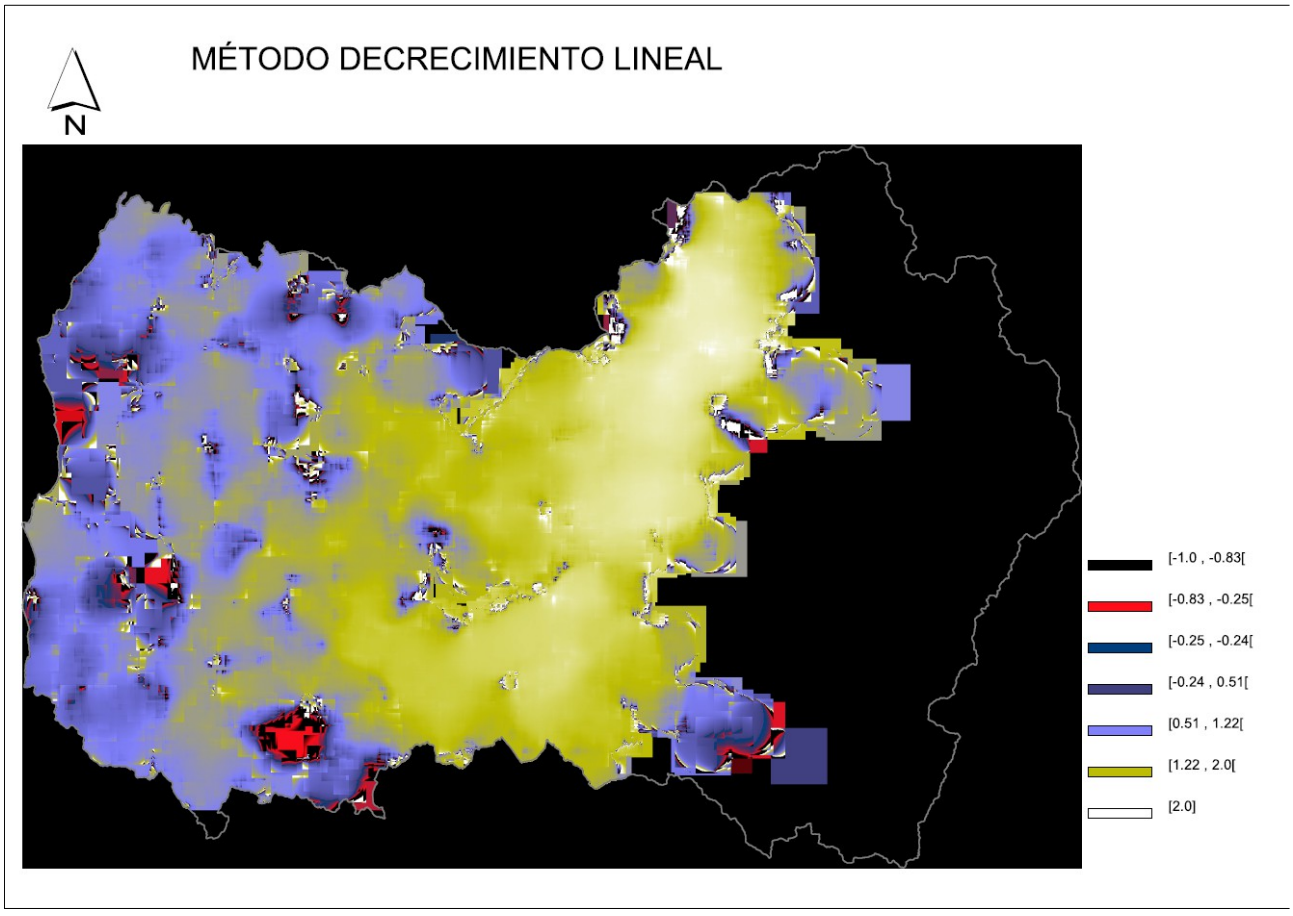

Ilustración 2

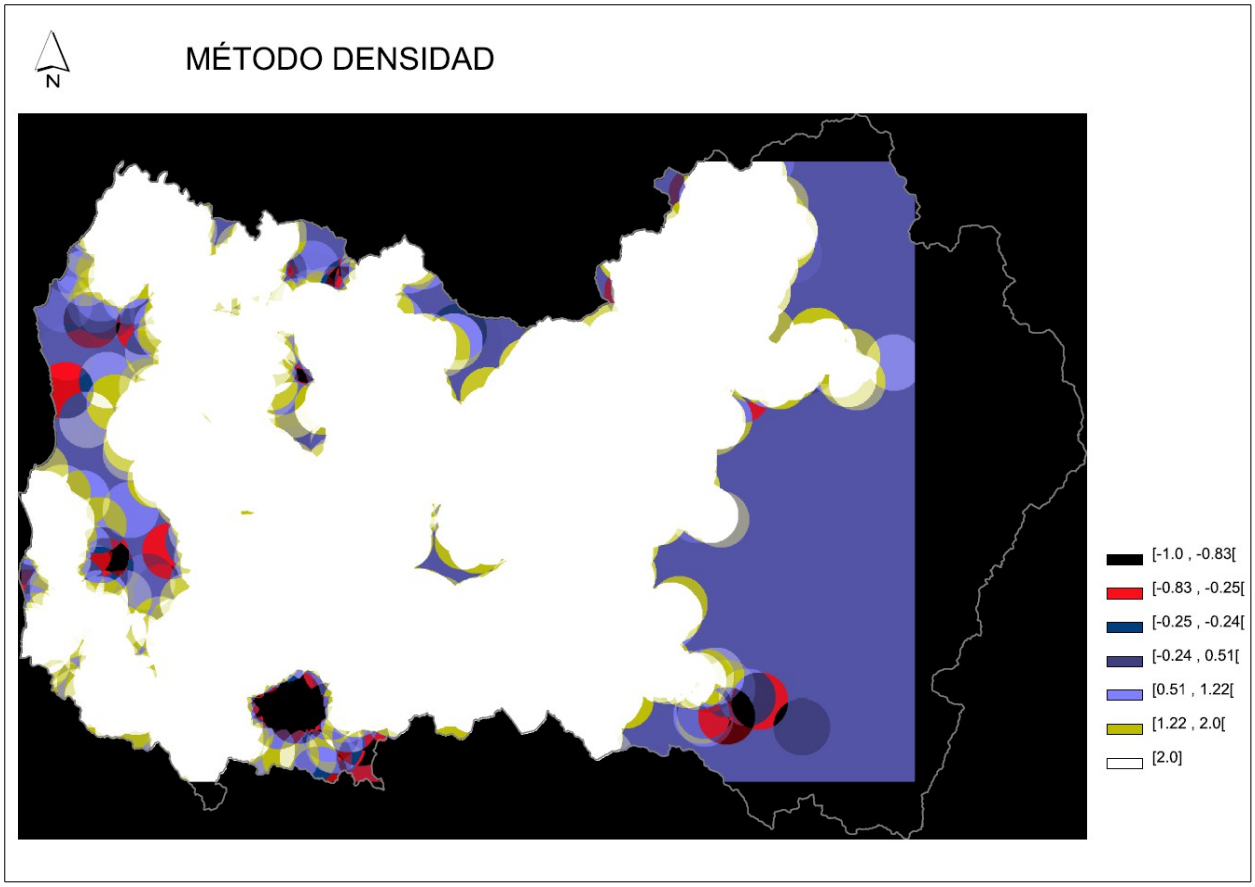

Ilustración 3

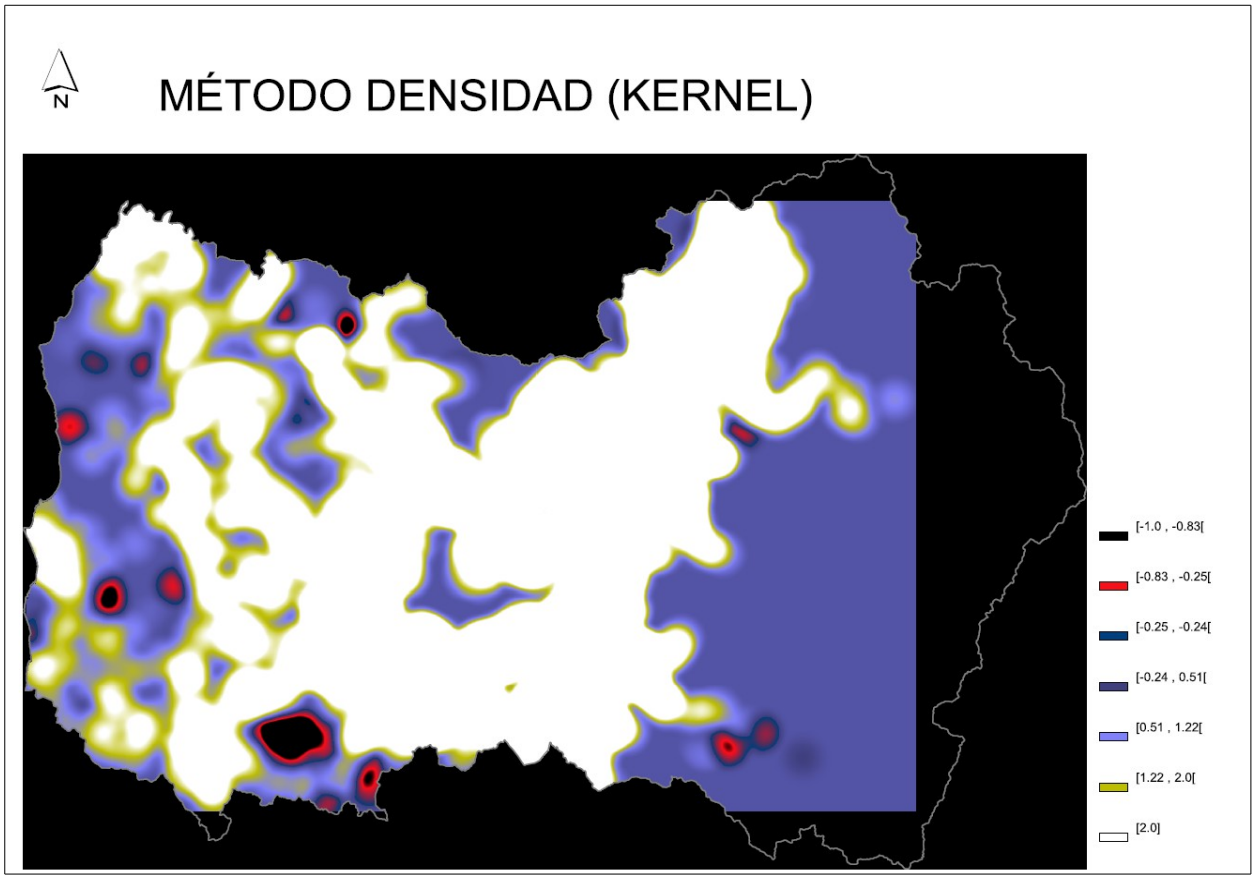

Ilustración 4

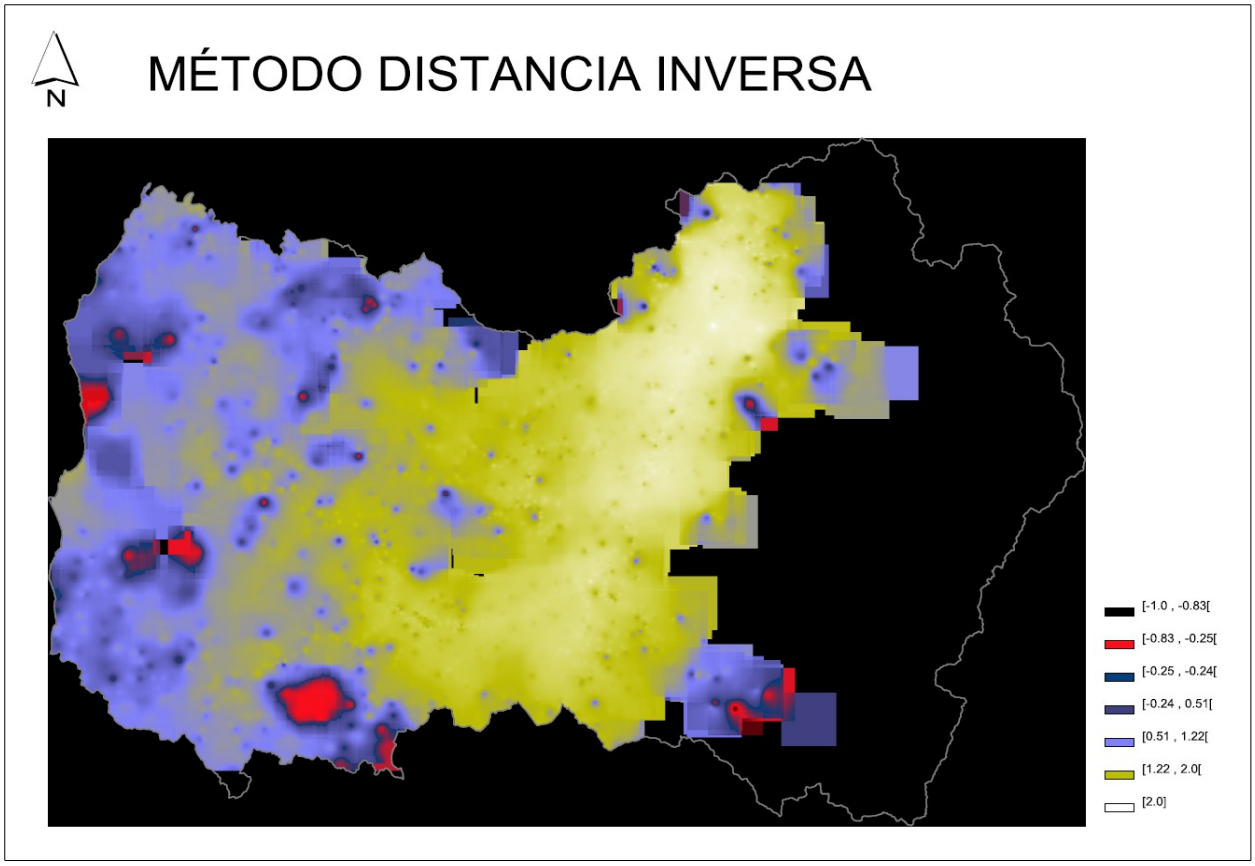

Ilustración 5

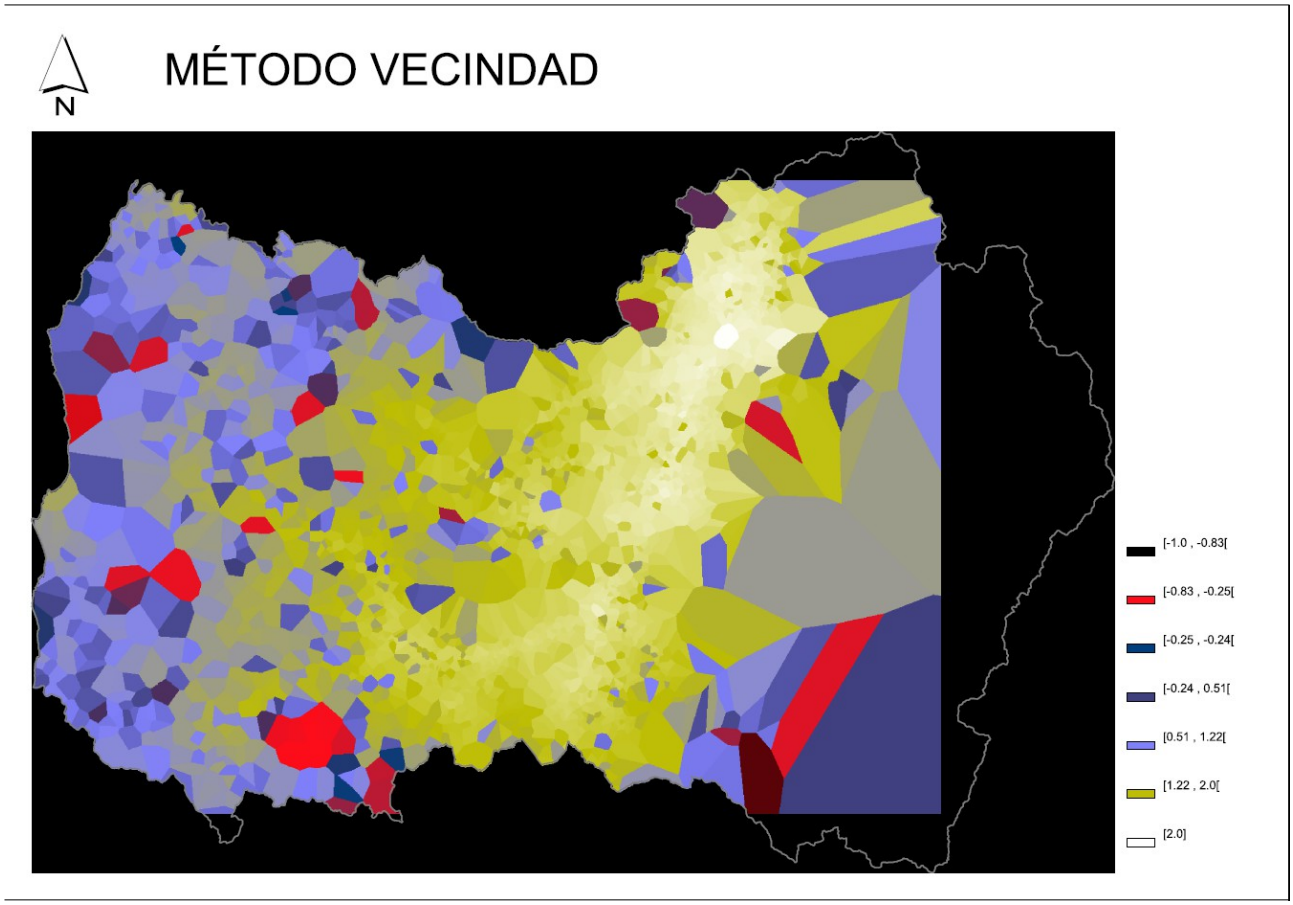

Ilustración 6

4.3. Medidas de ajuste.

A los 339 puntos de control, se le aplica el geoproceso muestreo de capas raster, con el fin de obtener el valor que toma el modelo raster en cada uno de esos puntos.

De la muestra de 339 localidades tomadas aleatoriamente, para ser puntos de control se procede a eliminar datos anómalos, por condiciones de borde (de tipo geográfico) los algoritmos de interpolación arrojaron datos erróneos o no calculados. Estos datos son tipificados con valores de **-10000000000, -99999, -96088.62846, -14042.82331, -593.39483.** Valores que se escapan del máximo y mínimo teórico del índice (2 a-1). Estos valores anómalos se explican debido a que los puntos de control quedaron en condición de borde, es decir, en esas coordenadas la interpolación fue deficiente debido a que no existían mas puntos para interpolar.

Eliminados estos valores la muestra queda en 334 localidades como puntos de control, arrojando los siguientes resultados Tabla 1

	R^2	A	MABSERR	MABSERR*
Decrecimiento Lineal	0,762671291	0,002249778	0,000418129	0,002923542
Densidad	-13483,57701	39,77473651	0,130127042	7,651236206
Densidad Kernel	-1858,412587	5,482066881	0,046504533	1,052710418
Distancia Inversa	0,817117003	0,002410385	0,000349931	0,002923905
Vecindad	0,726043418	0,00214173	0,00038594	0

Tabla 1

De los modelos revisados el que presenta mejores resultados en todas las medidas de ajustes es el método de la distancia inversa Ilustración 7 . En efecto, posee los mejores resultados en R^2 y MABSERR y para las medidas de ajuste de A y MABSRR* se observan valores próximos a los valores mas bajos obtenidos por otros métodos de interpolación.

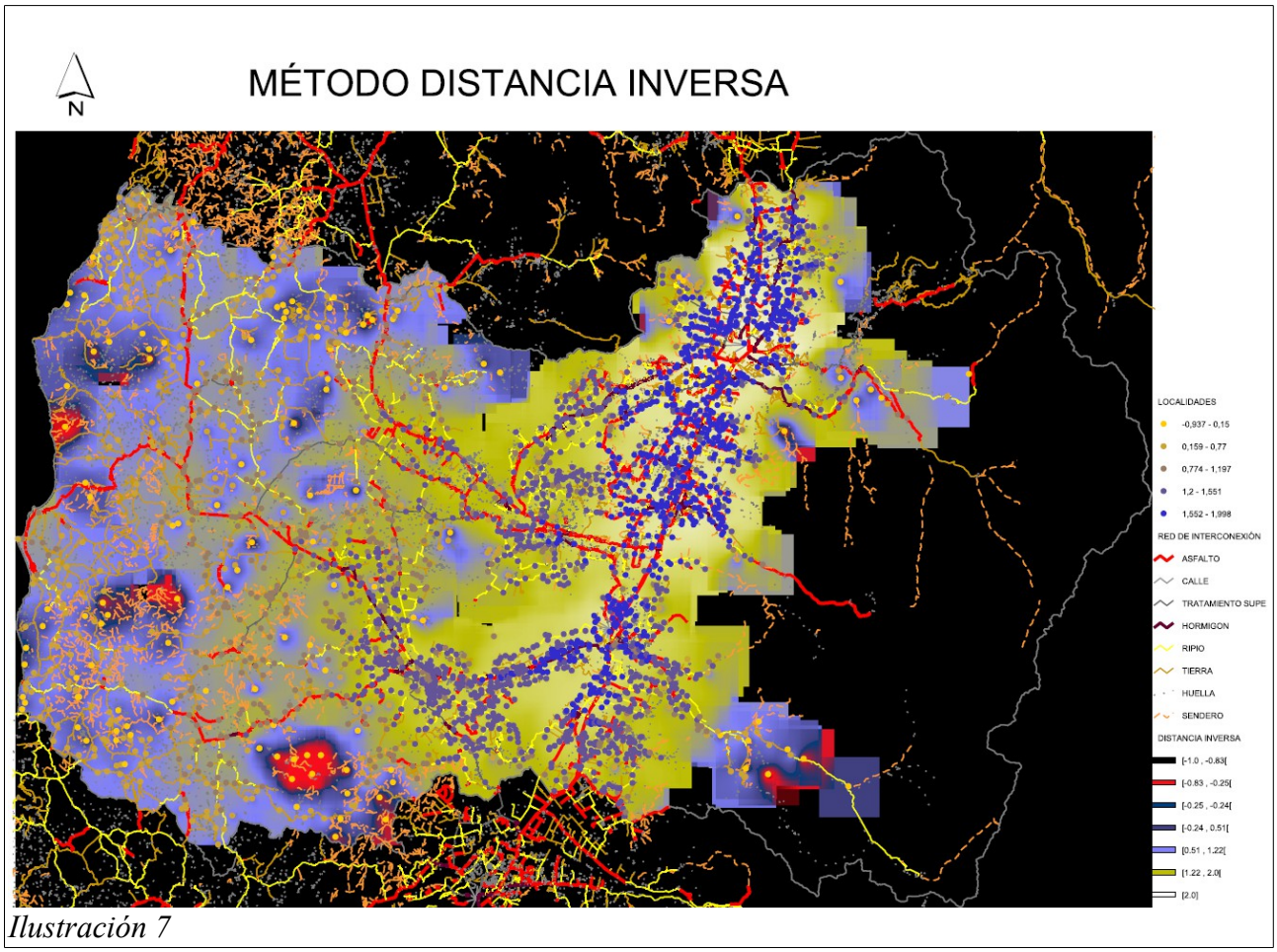

Ilustración 7

5. CONCLUSIONES.

Con respecto a software gvSIG.

Tiene un buen comportamiento en el proceso de interpolación, sin embargo, el proceso de muestreo con los puntos de control, no se puede realizar de forma simultánea con los 5 modelos raster (aun cuando tiene la opción para realizarlo).

La visualización con paleta de colores presenta algunos imperfectos, hay que realizar un zoom out para ver el despliegue a colores, ya que con zoom extendido solo se muestra un color.

Cabe destacar el carácter gratuito de este software, lo cual lo hace atractivo para realizar docencia.

Con respecto a los modelos raster generados a partir de métodos de interpolación.

A partir de las medidas de bondad de ajuste el método de interpolación más adecuado sería el de la distancia inversa, sin embargo, debe tener en cuenta que estas medidas de ajuste son válidas para la distribución de localidades de la región de O'Higgins y para los parámetros especificados. Por lo tanto, cambiando parámetros de los métodos de interpolación podrían generarse cambios en cada una de las medidas de bondad de ajuste.

Para determinar qué método de interpolación es el mejor, es importante conocer la naturaleza del fenómeno que se está estudiando. En este caso, el índice de aislamiento está calculado en función del tiempo (medido en horas) desde cada una de las localidades a una serie de servicios localizados en el territorio. La medición (tiempos) se realiza a través del cálculo de ruta mínima de la red de transporte, para esta región caminos de distintas categorías (hormigón, asfalto, tierra, ripio, huellas y senderos) . Esto es importante, ya que los métodos de interpolación, si bien consideran distancia mediante distintas funciones, esta es, la cartesiana o euclidiana.

Queda en evidencia que en las zonas donde exista una mayor "densidad" de caminos y localidades la interpolación será mejor, ya que existe un menor efecto de la sinuosidad[1] de los caminos. El fenómeno de aislamiento, que por definición esta localizado en lugares extremos, de difícil acceso y disperso, hace que los métodos en esas zonas tengan una sobre o sub estimación. Por lo que no es recomendable generar modelos de datos raster que representen este fenómeno a partir de métodos de interpolación. Sin perjuicio de lo anterior, sí se pueden generar buenas aproximaciones espaciales del fenómeno de integración, ya que como se mencionó, estas localidades estarían en lugares con una buena conexión, debido a la mayor densidad de vías de comunicación.

Consideraciones finales.

Las medidas de ajustes expuestas en este trabajo son de gran utilidad ya que permiten muestrear estadísticamente y validar, no solo métodos de interpolación, si no que cualquier salida raster proveniente de algún modelo (dispersión de contaminantes atmosféricos, suelo, agua, ruidos etc.), permitiendo realizar ajustes en los parámetros de calibración de ellos.

A pesar de su utilidad, es importante conocer el comportamiento del fenómeno a representar, dado que son métodos de validación de naturaleza estadística. A diferencia de realizar análisis visuales, donde se corre el riesgo de que la percepción de nuestro ojo, nos pueda llevar a realizar conclusiones erradas.

La elección del índice de aislamiento calculado para las localidades de la región de O'Higgins, fue ideada originalmente con el fin de contrastar las medidas de ajuste con métodos de interpolación que no representaban a cabalidad el fenómeno de aislamiento. Si se hubiesen tomado en cuenta solo las medidas de ajuste, obviando el conocimiento del fenómeno a estudiar, la conclusión hubiese sido errada ya que ningún método de interpolación (de los estudiados) toma el tiempo de desplazamiento por red vial.

1Distancia lineal entre dos puntos vs distancia "real"

6. BIBLIOGRAFIA.

- Hines W, Montgomery D, Goldsman D, Borror C. Probabilidad y Estadísticas para Ingeniería, 4ta ed.
- Manual gvSIG 1.11 http://www.gvsig.org/web/projects/gvsig-desktop/official/gvsig-1.11/descargas
- Miñarro A (1998). Estimación no paramétrica de la función densidad, Universidad de Barcelona.
- Vizuete D (2013). Determinación de los lugares de mayor incidencia de delitos y violencia en el Distrito de Quito con base en técnicas estadísticas espaciales.