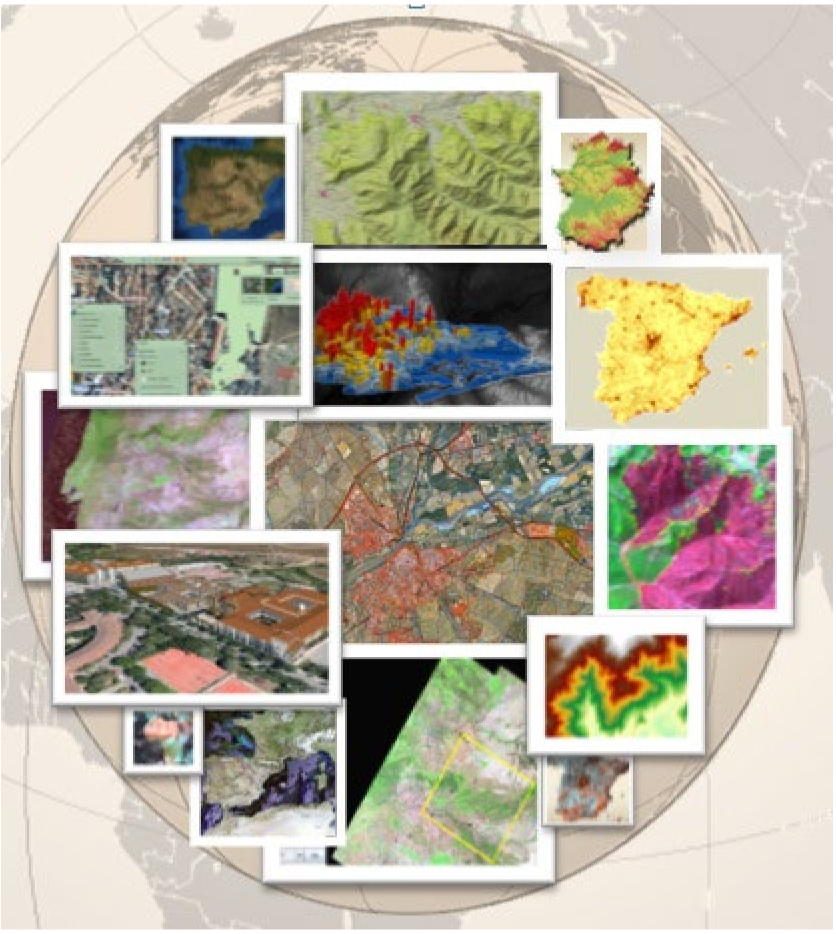
Máster en Tecnologías de la Información Geográfica: SIG y Teledetección



Geoestadística y calidad de la información

Tema 4. Aplicaciones estadísticas a modelos vectoriales

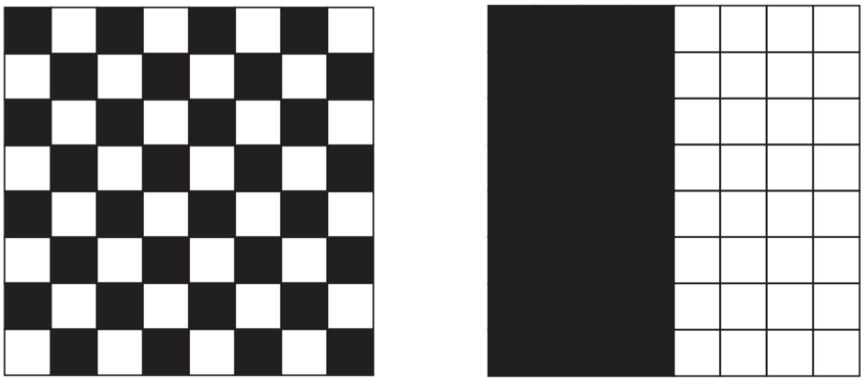
Dra. Elia Quirós Rosado

Contenido

1. Introducción

Una preocupación clave en el análisis de datos espaciales es examinar el patrón espacial en la variable o variables de interés. Por ejemplo, ¿son los valores de una determinada variable grandes en algunas zonas y pequeños en otras? Asimismo, ¿los valores similares tienden a agruparse o son visualmente erráticos? El término "dependencia espacial" se refiere a la dependencia de los valores vecinos entre sí.

La dependencia espacial se basa en que los valores cercanos tienden a parecerse más que los alejados. En el contexto de la medición estadística, esta idea está relacionada con la autocorrelación espacial, es decir, el grado en que una variable está correlacionada espacialmente consigo misma. Una medida de autocorrelación espacial puede sugerir dependencia espacial (es decir, los valores vecinos son similares -autocorrelación espacial positiva en la figura de la derecha-) o independencia espacial (los valores vecinos son disímiles -autocorrelación espacial negativa en la figura de la izquierda-).



Fuente: spatial data analysis an introduction for GIS users.

2. Índices de correlación espacial

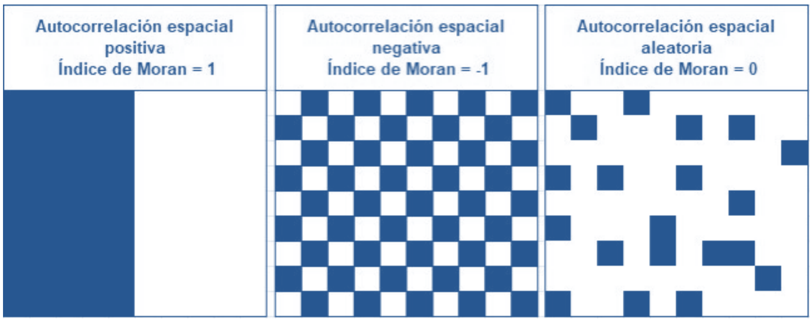
Los métodos de autocorrelación espacial más conocidos son I de Moran y G general de Getis-Ord. El último método tiene un carácter local, mientras que el primero permite analizar el espacio de una forma global.

2.1 Índice de Moran

Existen diversas medidas de autocorrelación espacial. Sin embargo, la medida de autocorrelación espacial que se encuentra más frecuentemente en la literatura del análisis espacial es el coeficiente I propuesto por Moran en 1950. Viene dado por:

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

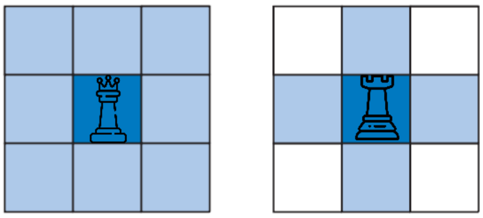
donde los valores y_i (de los cuales hay n) tienen la media $\bar{y}$ y la proximidad entre i y j viene dada por w_{ij}, que es un peso geográfico que a menudo se establece en 1 cuando las ubicaciones i y j son vecinas y en 0 cuando no lo son.



Los valores negativos de I indican una autocorrelación espacial negativa: los valores vecinos tienden a ser diferentes. Los valores positivos de I indican una autocorrelación espacial positiva: los valores vecinos tienden a ser similares. Los valores de I próximos a cero indican que no hay estructura.

Las relaciones espaciales se pueden buscar por distancias, por vecinos próximos o por contigüidad (cuando hablamos de entidades tipo polígono).

En este último caso la contigüidad se puede buscar siguiendo distintas conexiones topológicas:



Para este ejemplo, se comparan las celdas que comparten una arista con otra celda y no las celdas que sólo comparten esquinas. Por analogía con el movimiento de las piezas en el ajedrez, esto se denomina contigüidad del caso de la torre. Si también se incluyen las casillas que comparten esquinas (es decir, casillas conectadas en diagonal), se denomina contigüidad del caso de la reina. En la figura se ilustra la contigüidad del caso de la torre y del caso de la reina. En el caso de zonas de forma irregular, también se puede utilizar la contigüidad del caso de la torre y la contigüidad del caso de la reina, incluyendo esta última las zonas que están conectadas sólo por vértices además de por aristas, mientras que la primera sólo incluye las zonas unidas por aristas.

Las constricciones que tiene el cálculo del índice son las siguientes:

- Los resultados no son confiables con menos de 30 entidades.
- Todas las entidades deben tener al menos un vecino.
- Ninguna entidad debe tener todas las otras entidades como vecinos.
- En caso de que los valores de la variable estén sesgados, las entidades debían tener aproximadamente ocho vecinos cada una.

2.2 G general de Getis-Ord

Mide el grado de concentración para valores altos o bajos.

Consiste en una relación entre la media ponderada de los valores de las ubicaciones vecinas y la suma de todos los valores, sin incluir el valor de la ubicación (x_i):

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j} x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j}, \ \forall j \neq i$$

La interpretación de los estadísticos de Getis-Ord es muy sencilla: un valor mayor que la media sugiere un conglomerado Alto o un punto caliente, un valor menor que la media, indica un conglomerado Bajo o un punto frío. A diferencia del estadístico Moran local, el enfoque Getis-Ord no tiene en cuenta los valores atípicos espaciales.

La hipótesis nula para el estadístico establece que *no existe agrupación espacial de los valores de la variable.*

El estadístico G General Getis-Ord es más adecuado cuando se tiene una distribución bastante uniforme de los valores y se buscan picos espaciales inesperados de valores altos. Sin embargo, cuando los valores altos y bajos se agrupan, tienden a anularse mutuamente.

2.3 Comparación I Moran/ G general Getis-Ord

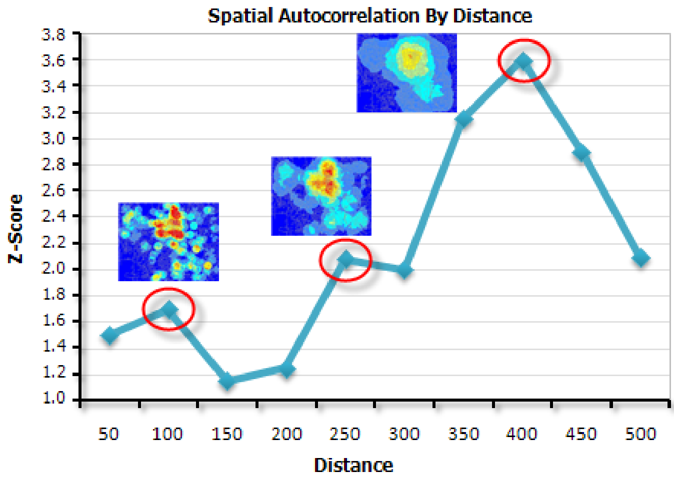
La hipótesis nula para el G General Getis-Ord y la autocorrelación espacial del índice de Moran es la aleatoriedad espacial completa; los valores se distribuyen aleatoriamente entre las características del conjunto de datos, reflejando procesos espaciales aleatorios.

Sin embargo, la interpretación del p-valor de cada una de ellas es diferente, como se muestra en la tabla siguiente:

<table>
<tr><th>Resultado</th><th>Agrupación alta/baja</th><th>Autocorrelación espacial</th></tr>
<tr><td>El p-valor no es estadísticamente significativo.</td><td colspan="2">No se puede rechazar la hipótesis nula. Es posible que la distribución espacial de los valores de los atributos de las características sea el resultado de procesos espaciales aleatorios. Dicho de otro modo, el patrón espacial de valores observado podría ser una de las muchas versiones posibles de una aleatoriedad espacial completa.</td></tr>
<tr><td>El p-valor es estadísticamente significativo y la puntuación z es positiva.</td><td>Se puede rechazar la hipótesis nula. La distribución espacial de los valores altos en el conjunto de datos está más agrupada espacialmente de lo que cabría esperar si los procesos espaciales subyacentes fueran realmente aleatorios.</td><td>Puede rechazar la hipótesis nula. La distribución espacial de valores altos y/o bajos en el conjunto de datos está más agrupada espacialmente de lo que cabría esperar si los procesos espaciales subyacentes fueran verdaderamente aleatorios.</td></tr>
<tr><td>El p-valor es estadísticamente significativo y la puntuación z es negativa.</td><td>Se puede rechazar la hipótesis nula. La distribución espacial de los valores bajos en el conjunto de datos está más agrupada espacialmente de lo que cabría esperar si los procesos espaciales subyacentes fueran realmente aleatorios.</td><td>Puede rechazar la hipótesis nula. La distribución espacial de valores altos y bajos en el conjunto de datos está más dispersa espacialmente de lo que cabría esperar si los procesos espaciales subyacentes fueran verdaderamente aleatorios. Un patrón espacial disperso suele reflejar algún tipo de proceso competitivo: un rasgo con un valor alto repele a otros rasgos con valores altos; del mismo modo, un rasgo con un valor bajo repele a otros rasgos con valores bajos.</td></tr>
</table>

2.4 Autocorrelación espacial incremental

Mide la autocorrelación espacial a una serie de distancias y, opcionalmente, crea un gráfico lineal de esas distancias y sus correspondientes puntuaciones z. Las puntuaciones z reflejan la intensidad de la agrupación espacial, y las puntuaciones z máximas estadísticamente significativas indican las distancias en las que los procesos espaciales que promueven la agrupación son más pronunciados. Estas distancias máximas suelen ser valores apropiados para herramientas con un parámetro de distancias de banda o radio de distancia.



3. Variables regionalizadas

La base de la geoestadística es la teoría de las variables regionalizadas. La geoestadística implica una división conceptual de la variación espacial (en un lugar *x*) en dos componentes distintas: una componente determinista (μ(x)) (que representa el cambio "gradual" en el área de estudio) y una componente estocástica (o "aleatorio") (R(x)):

$$Z(\boldsymbol{x}) = \mu(\boldsymbol{x}) + R(\boldsymbol{x})$$

Se trata de un modelo de "función aleatoria" (RF). La parte aleatoria refleja nuestra incertidumbre sobre las variables espaciales: lo que nos parece aleatorio es una función de una multiplicidad de factores que pueden ser imposibles de modelizar directamente (y esto no significa que realmente pensemos que la variación es aleatoria). En geoestadística, una variable referenciada espacialmente, Z(x), se trata como un resultado de una RF, Z(x). En otras palabras, consideramos efectivamente que una observación ha sido generada por el modelo de RF y esto nos proporciona un marco para trabajar con estos datos. Una realización de una RF se denomina variable regionalizada. La teoría de las variables regionalizadas es el marco fundamental en el que se basa la geoestadística.

En términos prácticos, al igual que estimamos los parámetros, es decir, la media y la varianza, de una distribución, estimamos los parámetros del modelo de RF utilizando los datos. Estos parámetros, al igual que la media y la varianza, resumen la variable. La media y la varianza de una de una distribución sólo son útiles si la distribución es aproximadamente normal y, del mismo modo, los parámetros del modelo de RF sólo son significativos en determinadas condiciones. Cuando las propiedades de la variable de interés son las mismas, o al menos similares en algún sentido, en toda la región de interés podemos emplear un modelo estacionario. En otras palabras, podemos utilizar los mismos parámetros del modelo en todos los lugares. Si las propiedades de la variable son claramente variables desde el punto de vista espacial, un modelo de RF estándar puede no ser adecuado. Existen diferentes grados de estacionariedad, pero en este tema sólo consideraremos uno, la **estacionariedad intrínseca**. Existen dos requisitos para la estacionariedad intrínseca:

- La media es constante en toda la región de interés. En otras palabras, el valor esperado de la variable no depende de la ubicación, *x*.
- La diferencia al cuadrado esperada entre RF emparejadas (es decir, las observaciones) debería depender únicamente de la distancia de separación y de la dirección entre los puntos y no de su ubicación.

A su vez, una variable regionalizada, es una variable distribuida en el espacio de forma que presenta una estructura espacial de correlación.

Podemos hacernos las siguientes preguntas:

¿Hay una correlación? ¿Cómo de sólida es la relación? ¿Qué variables son los indicadores más consistentes? ¿Las relaciones son consistentes en toda el área de estudio?

Para ello, existen diversos métodos geoestadísticos que dan respuestas a esas preguntas:

3.1 Mínimos cuadrados ordinarios (OLS)

La expresión general de un modelo de regresión, para un total de k variables explicativas es:

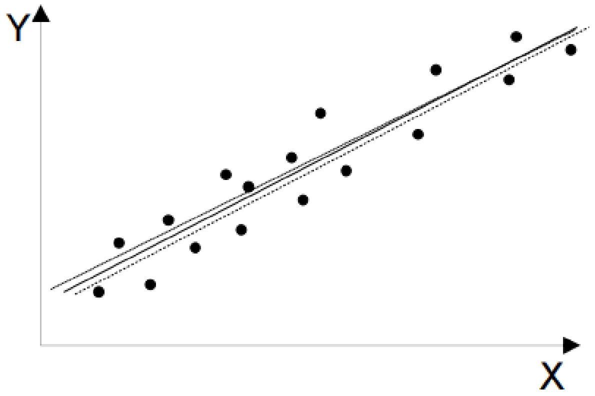
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + U$$

donde Y es la variable explicada, las X_j son las variables explicativas, y los parámetros β_j son unos parámetros que cuantifican la relación existente entre la variable explicada y cada variable explicativa. El término U, es la variable aleatoria que recoge la influencia sobre la variable explicada de otras variables explicativas que no hemos tenido en cuenta en el modelo. El término U es necesario ya que la variable explicada (Y) es una variable aleatoria, pero al otro lado de la igualdad (en el modelo), no son aleatorios ni los parámetros β ni las variables explicativas X. Como debe existir una parte aleatoria en ese lado del modelo, añadimos este término U.

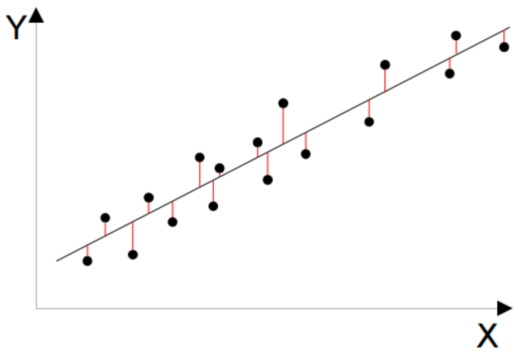
El error del modelo de regresión será la diferencia entre el valor real de la variable y el valor propuesto para la misma en el modelo. A la hora de generar el modelo, lo deseable es que el error fuera lo más pequeño posible. En cada punto de estudio conocemos el valor de todas las variables, luego tendremos tantas ecuaciones como puntos disponibles:

$$\begin{gathered} Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \cdots + \beta_k X_{k1} + U_1 \\ Y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \cdots + \beta_k X_{k2} + U_2 \\ \cdots \\ Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \cdots + \beta_k X_{kn} + U_n \end{gathered}$$

Para resolver un problema por mínimos cuadrados, debemos tener más ecuaciones que incógnitas. Por tanto, tendremos múltiples rectas o hiperplanos (según el número de variables explicativas) posibles para explicar la relación entre las variables:



El método de los mínimos cuadrados ordinarios consiste en la obtención de un hiperplano de forma que se minimice la suma de los cuadrados de las distancias entre cada una de las observaciones de la variable y dicho hiperplano (residuos).



Para ello no sólo se miden las distancias, sino que se estima el error sumando al cuadrado todas ellas para que las negativas no anulen a las positivas y al elevarlo al cuadrado, penalizamos mucho más los puntos que están más alejados, y que tienen más error, para que no ponderen mucho en la estimación del mejor hiperplano. De ahí viene el nombre de mínimos cuadrados.

3.2 Regresión exploratoria

Encontrar un modelo OLS correcto puede ser difícil, especialmente cuando hay muchas posibles variables explicativas que podrían ser importantes para modelar la variable dependiente. La Regresión exploratoria es una herramienta de extracción de datos que realiza todas las combinaciones posibles de variables explicativas para ver qué modelos superan todos los estadísticos de OLS necesarios. Al evaluar todas las combinaciones posibles de las posibles variables explicativas, aumentan las posibilidades de encontrar el mejor modelo.

Aunque la Regresión exploratoria es similar a la Regresión por pasos, en lugar de solo buscar modelos con valores altos de R^2 ajustada, la Regresión exploratoria busca modelos que cumplan con todos los requisitos y suposiciones del método de OLS.

Este tipo de estudios podrían ser rechazados por algunos investigadores ya que, en teoría, se deben formalizar las hipótesis antes de explorar los datos para evitar la creación de modelos que se ajusten solo a los datos concretos, y no reflejen los procesos más amplios. Podría ser el caso de que mediante esta técnica encontráramos modelos que se ajustaran muy bien a nuestros datos, pero no funcionaran con otros de las mismas características.

La mejor forma de actuar es seleccionar previamente variables de regresión explicativa compatibles con el sentido común. Después, construir los modelos de regresión con una parte de sus datos, por ejemplo, el 70% de los datos y validarlos con el resto (30%).

3.3 Regresión ponderada geográficamente (GWR)

La regresión ponderada geográficamente es una forma local de regresión lineal que se utiliza para modelar las relaciones entre la variable dependiente y explicativas que varían espacialmente. En el caso de los datos geográficamente dispersos con variabilidad espacial inherente, la estimación de los coeficientes de un modelo de regresión para una ubicación concreta basada únicamente en las observaciones de dicha ubicación no es factible debido al reducido número de observaciones. La regresión ponderada geográficamente es una herramienta importante para explorar la no estacionariedad espacial de la relación de regresión en el análisis de datos espaciales.

Mediante el uso de GWR, se construye una ecuación de regresión distinta para cada observación del conjunto de datos al incorporar las variables dependiente y explicativas de las entidades vecinas que caen dentro de un ancho de banda.

Al estimar el parámetro para una localización especifica, los elementos en los datos se ponderan según su distancia a esta localización, con mayor peso para los elementos más cercanos.

El modelo GWR puede ser descrito de la siguiente forma:

$$y(s) = \beta_1(s)x_1(s) + \ldots + \beta_p(s)x_p(s) + \epsilon(s)$$

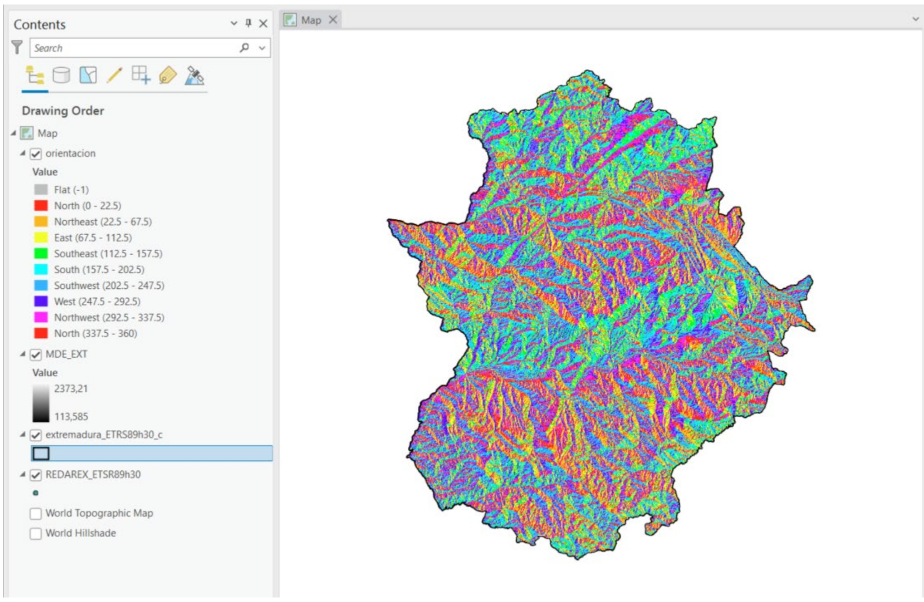
donde *y(s)* es la variable de respuesta en el lugar *s*, $\beta_i(s)$, $i = 1, 2, \ldots, p$ son los coeficientes de las variables independientes en el lugar s, y ϵ(s) es el efecto aleatorio en el lugar s también llamado "intercepción".

El tipo de modelo puede ser: Gaussiano, Logístico o Poisson, a diferencia del OLS, que siempre es lineal. Cada uno de ellos se corresponde con variables continuas (por ejemplo, valores de temperatura), binarias (por ejemplo, presencia o ausencia) o de recuento (por ejemplo, número de ocurrencias de un suceso).

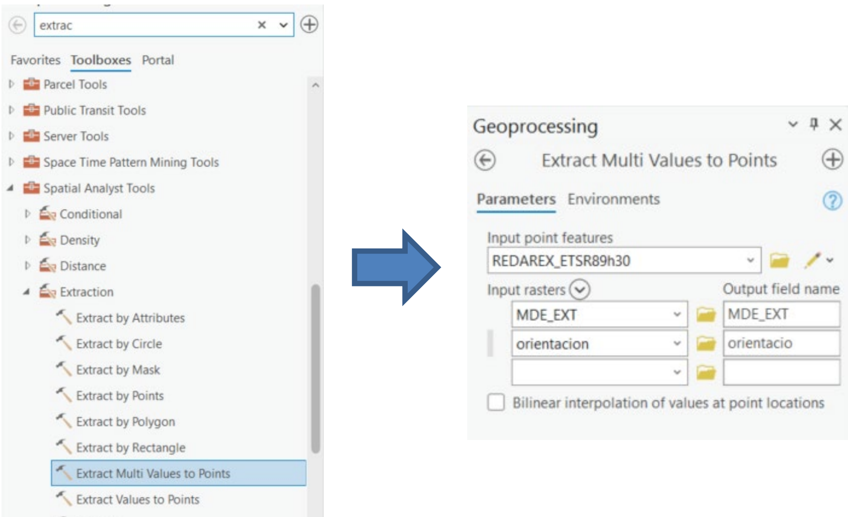
Las relaciones de vecindad, como en otros estadísticos, pueden darse o bien por distancia o valor máximo de vecinos y con respecto al peso de los elementos más cercanos se puede usar un esquema de ponderación Gaussiano o Bicuadrado. El primero disminuye gradualmente a medida que aumenta la distancia de la entidad de regresión. El segundo es igual, pero todas las entidades que quedan fuera de la vecindad especificada se les asigna cero y no afectan a la regresión local.

4. Análisis y modelización espacial con ArcGIS

En primer lugar, cargaremos de nuevo todas las estaciones de Redarex. Cargaremos también las dos capas del MDE y la orientación de las laderas de Extremadura.



Vamos a extraer en cada estación, el valor de la elevación del MDE y la orientación. Para ello, emplearemos la herramienta *extract multivalues to points*:



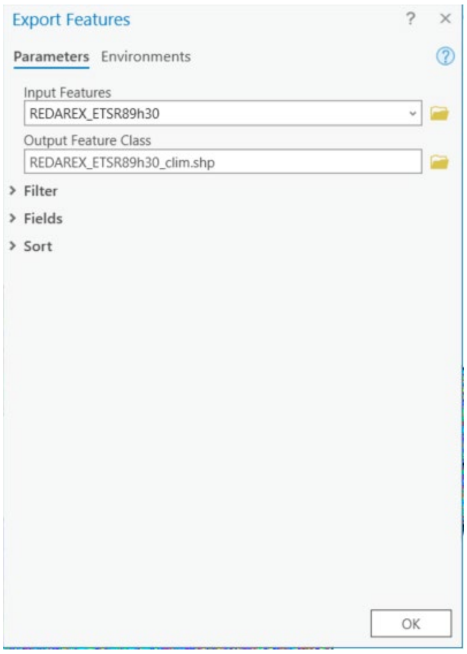
Ahora comprobaremos que se ha añadido esa información a la tabla de atributos de los puntos.

	ion	Code	X_web	Y_web	layer	path	MDE_EXT	orientacio
1	ıtara (CC104)	CC104	680082	4401750	Est_REDAREX_web_h29...	F:\Precipitation\Precipit...	333,512	166,606
2	ıuela del Jerte (CC...	CC04	736412	4432357	Est_REDAREX_web_h29...	F:\Precipitation\Precipit...	262,05	291,058
3	da (CC106)	CC106	704113	4349955	Est_REDAREX_web_h29...	F:\Precipitation\Precipit...	315,403	105,002
4	yo (BA206)	BA206	718134	4304619	Est_REDAREX_web_h29...	F:\Precipitation\Precipit...	212,396	39,5195
5	ga (BA102)	BA102	263512	4252720	Est_REDAREX_web_h30	F:\Precipitation\Precipit...	542,165	298,501
6	al (BA203)	BA203	688422	4305380	Est_REDAREX_web_h29...	F:\Precipitation\Precipit...	186,212	358,981
7	tejada (CC01)	CC01	270938	4416527	Est_REDAREX_web_h30	F:\Precipitation\Precipit...	269,092	237,119
8	ı-Puebla de Argem...	CC14	204857	4429511	Est_REDAREX_web_h30	F:\Precipitation\Precipit...	236,686	231,867
9	Benito-EFA (BA105)	BA105	252468	4319179	Est_REDAREX_web_h30	F:\Precipitation\Precipit...	253,072	338,461

Y finalmente volveremos a hacer la unión (del mismo modo que en el tema 3) con los datos climáticos calculados en la primera parte de la asignatura.

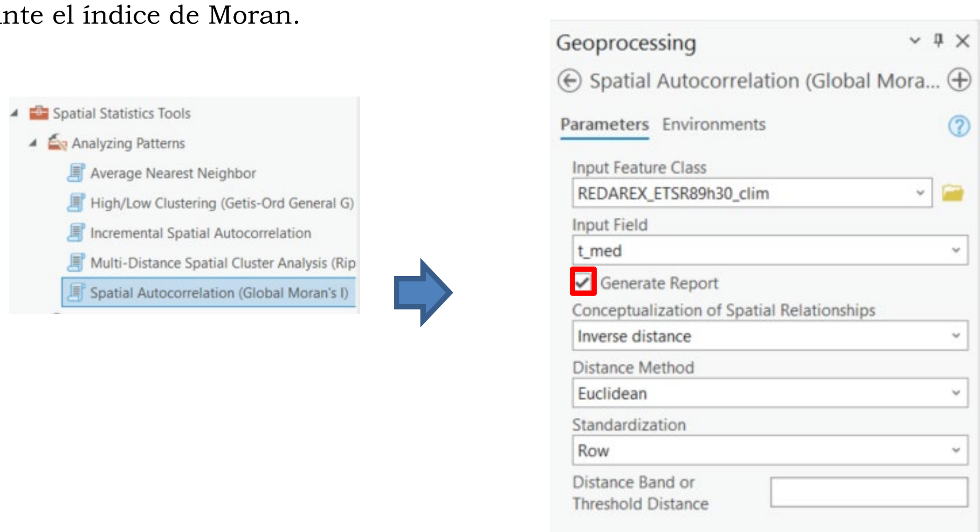
		MDE_EXT	orientacio	F1	Estación	Identificador	t_max	t_med	t_min	preci	oscial_ter
29	ipitation\Precipit...	445,618	41,3901	1	Zalamea de la Serena	BA01	24,49	18,07	11,68	307,07	12,
30	ipitation\Precipit...	264,83	204,136	28	Moraleja	CC16	24,08	17,12	10,19	634,96	13,
31	ipitation\Precipit...	495,013	18,1597	2	Monterrubio de la Serena	BA02	24,24	17,58	10,99	357,41	13,
32	ipitation\Precipit...	503,129	131,043	10	Puebla de Alcocer	BA103	23,6	18,13	12,83	629,69	10,
33	ipitation\Precipit...	333,512	166,606	34	Alcantara	CC104	23,88	17,87	12,12	545,64	11,
34	ipitation\Precipit...	542,971	150,528	35	Mirabel	CC105	22,04	16,98	12,09	794,88	9,
35	ipitation\Precipit...	568,012	343,958	29	Valdastillas	CC17	22,97	17	11,86	1119,74	11,
36	ipitation\Precipit...	763,318	117,554	33	Guadalupe	CC103	20,96	16,94	12,96	973,98	

Consolidaremos la unión guardando un nuevo shp con todos los datos a modelizar.

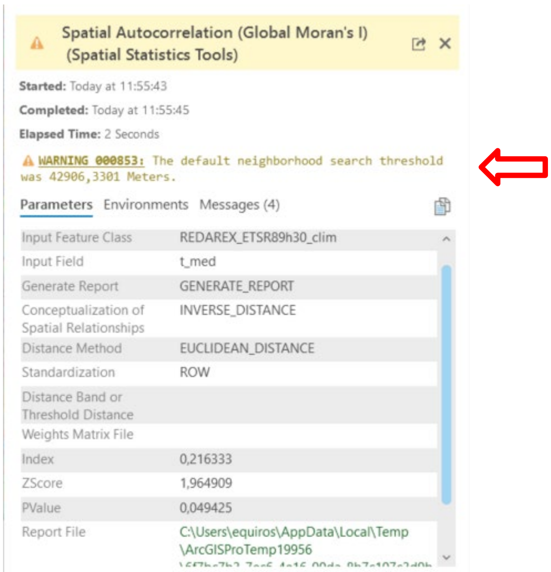


4.1 Índice de Moran

Para calcular este índice, abriremos la caja de estadísticas espaciales y dentro del cajón de análisis de patrones encontraremos el cálculo de la autocorrelación espacial mediante el índice de Moran.

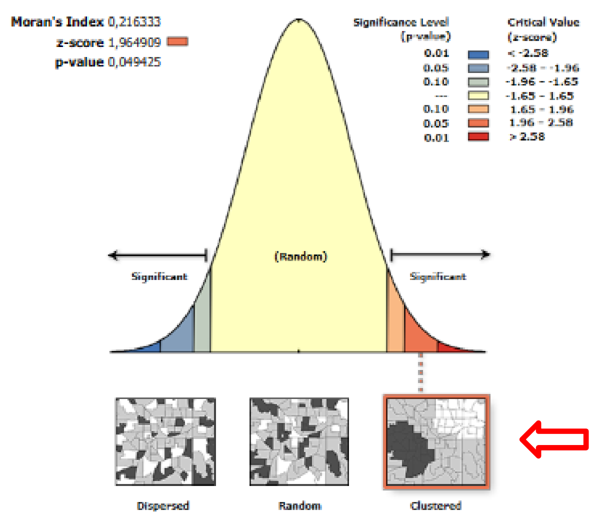


Se nos abrirá una ventana emergente en la que nos advertirá inicialmente de que ha utilizado la mínima distancia para hacer el cálculo y en la parte final, nos indicará la ruta en la que se encuentra el informe de autocorrelación:



Nota, hemos realizado el caso más sencillo de conceptualización de relaciones espaciales que es el de la inversa de la distancia, en el caso de que hubiéramos elegido el de la contigüidad de polígonos, habríamos podido elegir entre el movimiento de la reina o de la torre explicado anteriormente.

Spatial Autocorrelation Report



Given the z-score of 1.964909, there is a less than 5% likelihood that this clustered pattern could be the result of random chance.

Global Moran's I Summary

Moran's Index	0,216333
Expected Index	-0,028571
Variance	0,015535
z-score	1,964909
p-value	0,049425

Dataset Information

Input Feature Class:	REDAREX_ETSR89h30_clim
Input Field:	T_MED
Conceptualization:	INVERSE_DISTANCE
Distance Method:	EUCLIDEAN
Row Standardization:	True
Distance Threshold:	42906,3301 Meters
Weights Matrix File:	None
Selection Set:	False

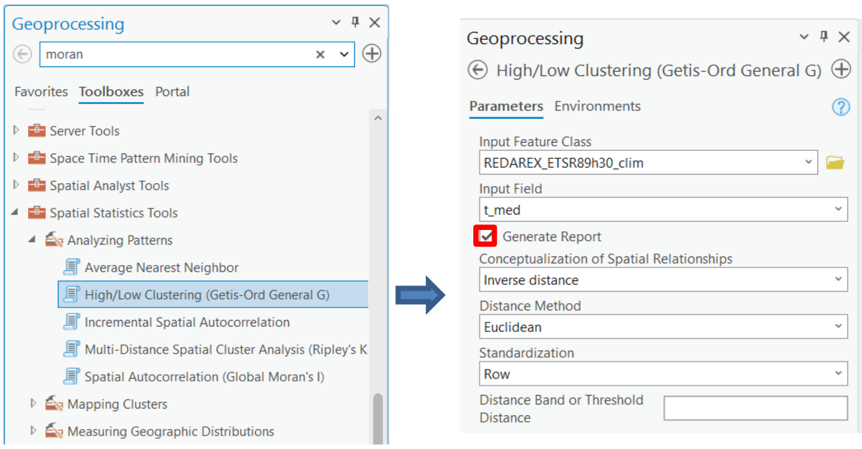
La herramienta Autocorrelación espacial devuelve cinco valores: el índice I de Moran, el índice esperado, la varianza, la puntuación z y el valor p.

Un valor positivo del índice I de Moran indica tendencia hacia la agrupación, mientras que un valor negativo del índice I de Moran indica tendencia hacia la dispersión. El índice de Moran obtenido es de 0.21633, luego podemos asegurar un agrupamiento de las temperaturas medias.

La puntuación z y el valor p son medidas de significación estadística que le indican si debe o no rechazar la hipótesis nula. En este caso la hipótesis nula establece que “los valores de temperatura media se distribuyen aleatoriamente”. Como el valor de puntuación Z es 1.964909, hay menos de un 5% de probabilidades de que este patrón agrupado pueda ser el resultado del azar.

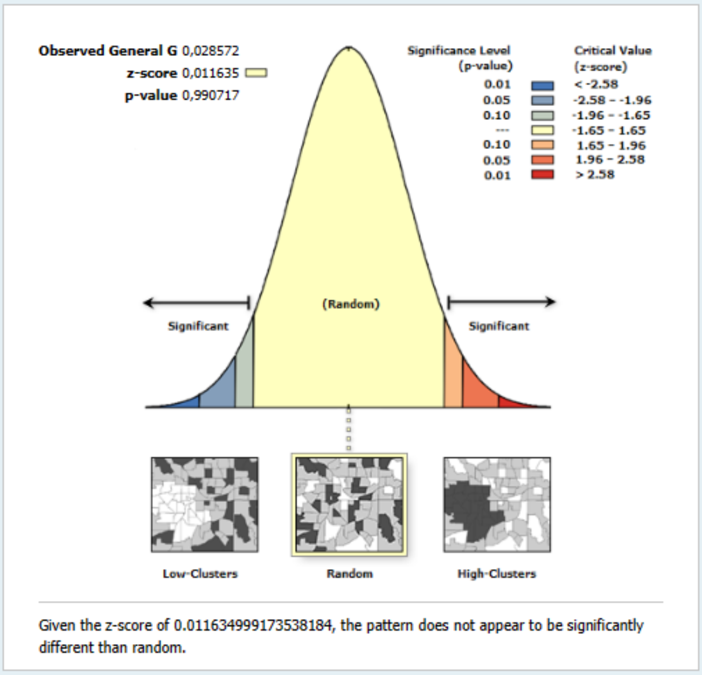
4.2 G general de Getis-Ord

En este caso iremos a la herramienta correspondiente y la configuraremos de la siguiente manera:



Después del mismo aviso que en el I. de Moran acerca de la distancia, nos aparece el enlace donde consultar los resultados del cálculo:

High-Low Clustering Report



General G Summary

Observed General G	0,028572
Expected General G	0,028571
Variance	0,000000
z-score	0,011635
p-value	0,990717

Dataset Information

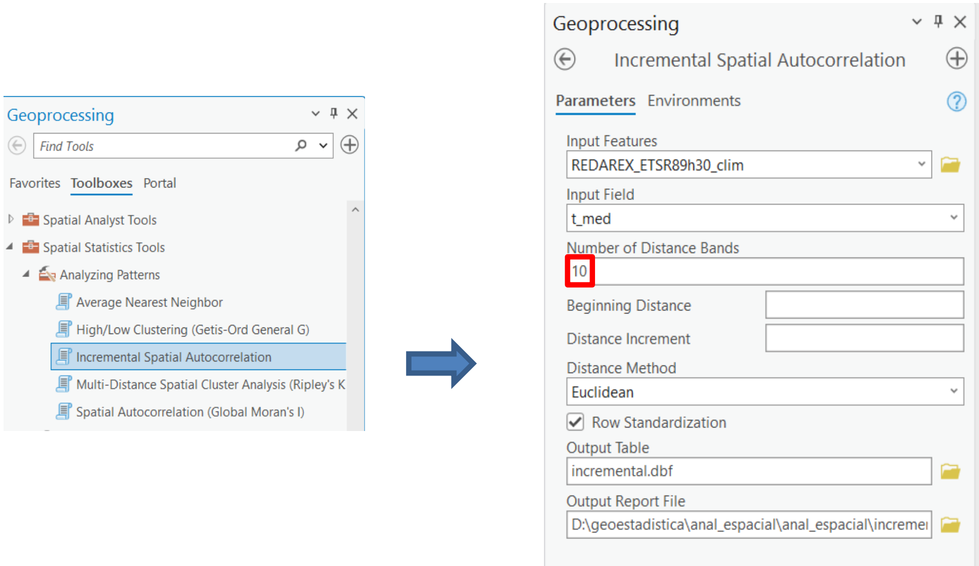
Input Feature Class:	REDAREX_ETSR89h30_clim
Input Field:	T_MED
Conceptualization:	INVERSE_DISTANCE
Distance Method:	EUCLIDEAN
Row Standardization:	True
Distance Threshold:	42906,3301 Meters
Weights Matrix File:	None
Selection Set:	False

Cuando el valor p devuelto por esta herramienta es pequeño y estadísticamente significativo, puede rechazarse la hipótesis nula. Si se rechaza la hipótesis nula, el signo de la puntuación z adquiere importancia. Si el valor de la puntuación z es positivo, el índice General G observado es mayor que el índice General G esperado, lo que indica que los valores altos del atributo están agrupados en la zona de estudio. Si el valor de la puntuación z es negativo, el índice General G observado es menor que el índice esperado, lo que indica que los valores bajos están agrupados en el área de estudio.

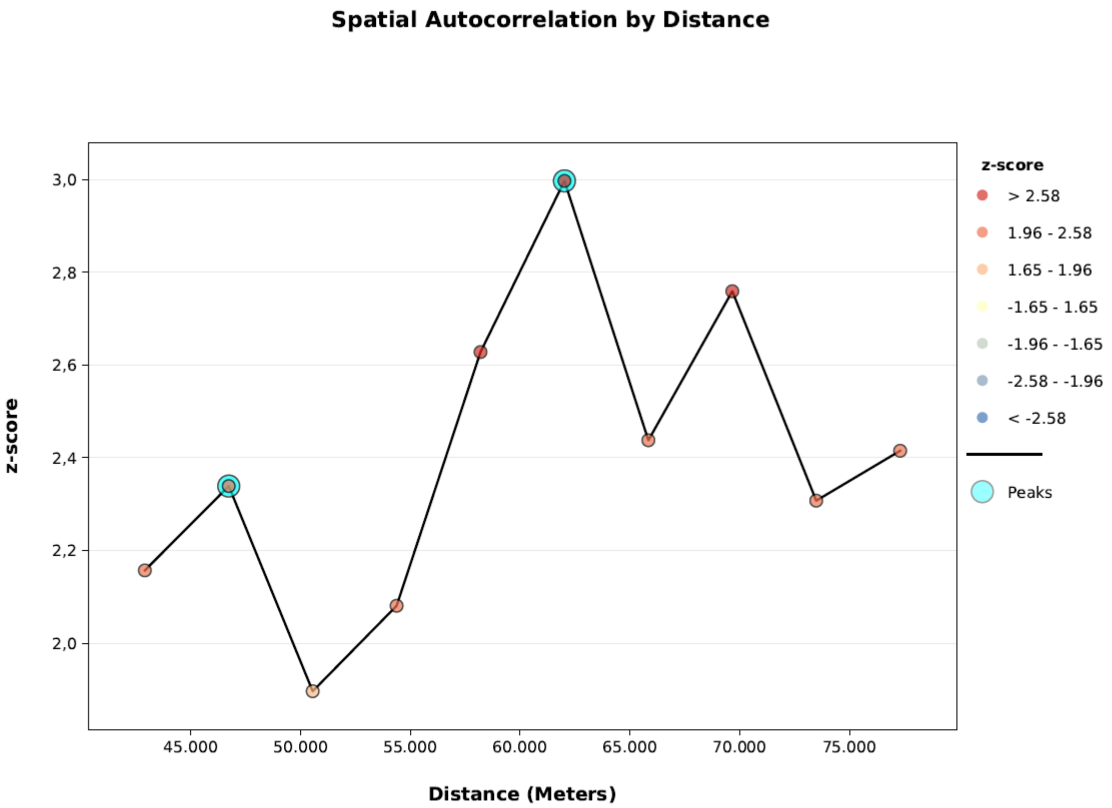
<u>En nuestro caso, el p-valor es alto, por lo que no se puede rechazar la hipótesis nula y podemos decir que los valores altos-bajos están aleatoriamente distribuidos.</u>

4.3 Autocorrelación espacial incremental

Localizaremos la herramienta y la configuraremos de la siguiente forma:



Una vez calculado podremos abrir un pdf en el que se nos muestran los resultados.



Podemos observar que el primer pico es el que se encuentra a 46725 m y tiene una puntuación z de 2,3389. El segundo está a 62017 m y tiene una puntuación z de 2,9972. En ambos casos, los valores de los estadísticos indican agrupamiento espacial de las temperaturas medias.

Global Moran's I Summary by Distance

Distance	Moran's Index	Expected Index	Variance	z-score	p-value
42903,00	0,230893	-0,028571	0,014469	2,157019	0,031004
46725,81	0,226474	-0,028571	0,011890	2,338954	0,019338
50548,62	0,154490	-0,028571	0,009318	1,896433	0,057903
54371,44	0,163930	-0,028571	0,008559	2,080711	0,037460
58194,25	0,183083	-0,028571	0,006486	2,628162	0,008585
62017,06	0,192231	-0,028571	0,005427	2,997280	0,002724
65839,87	0,136912	-0,028571	0,004608	2,437783	0,014778
69662,68	0,147221	-0,028571	0,004059	2,759257	0,005793
73485,50	0,104590	-0,028571	0,003330	2,307478	0,021028
77308,31	0,103997	-0,028571	0,003013	2,415121	0,015730

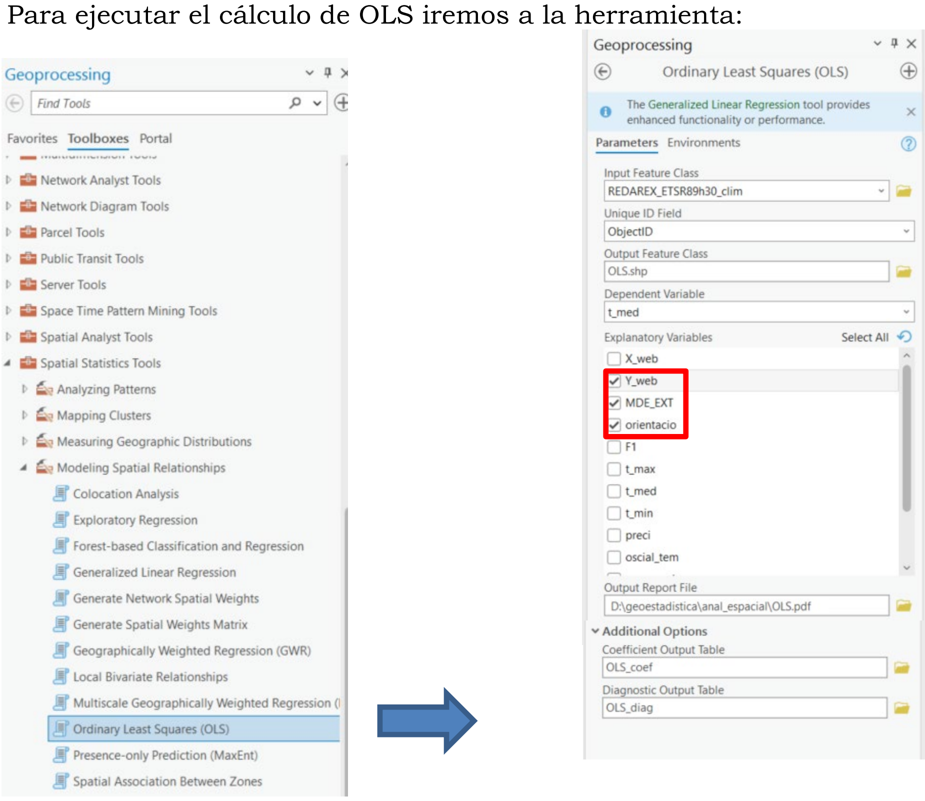
First Peak (Distance; Value): 46725,81; 2,338954

Max Peak (Distance; Value): 62017,06; 2,997280

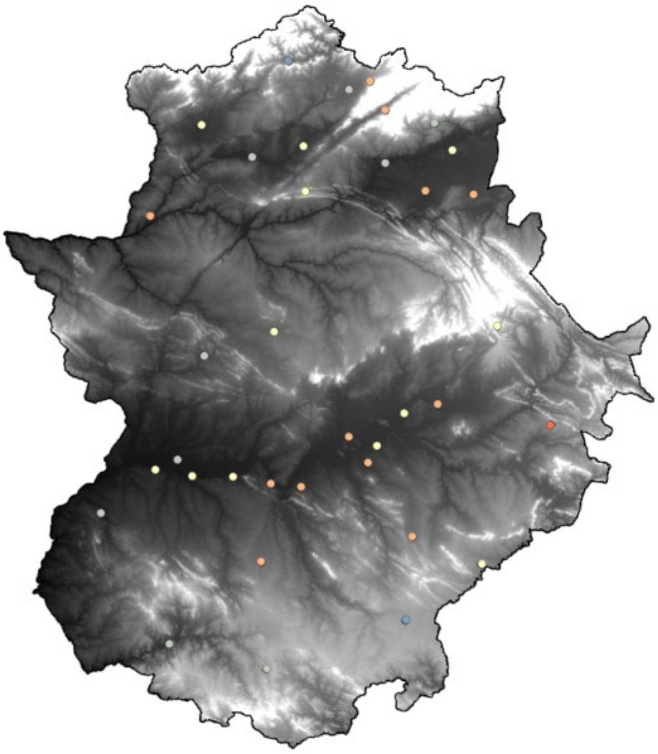
Distance measured in Meters

4.4 Mínimos cuadrados ordinarios

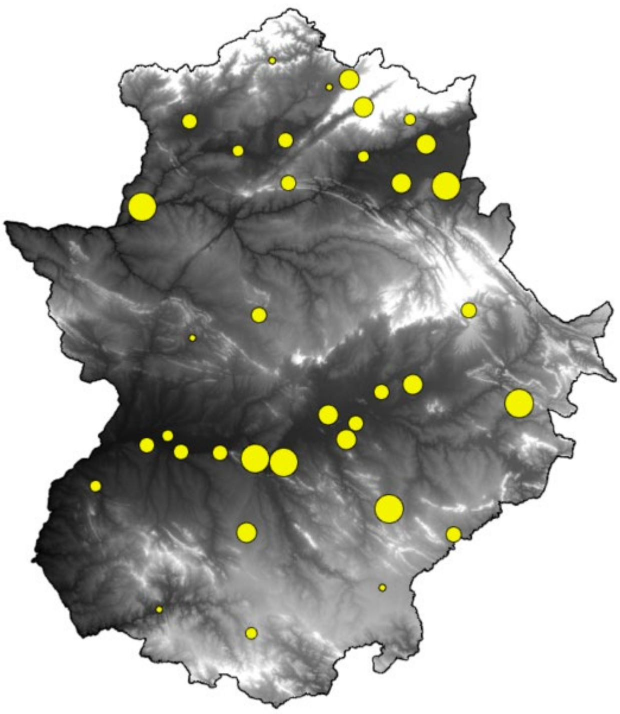
Para ejecutar el cálculo de OLS iremos a la herramienta:

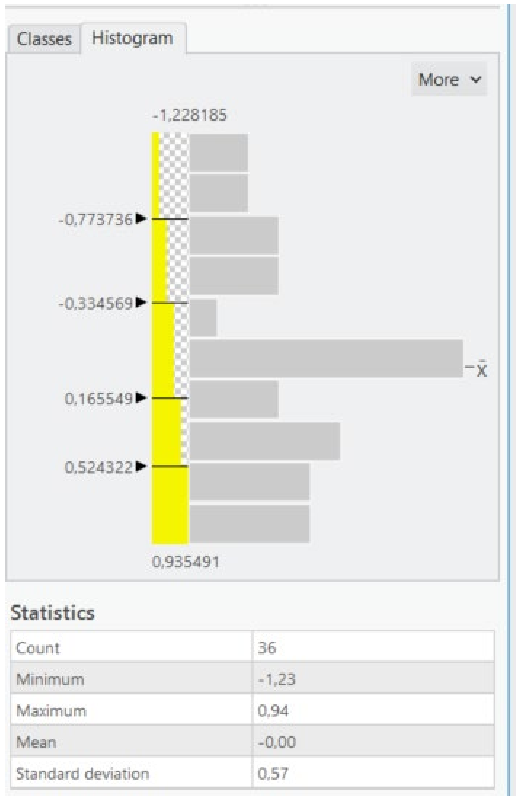


Se nos habrá generado un nuevo shp con los errores de cada punto con respecto al mejor modelo explicativo.



Para visualizar mejor los puntos con mayor error vamos a representar los puntos por tamaños (el tamaño en función del residuo):





Si analizamos el informe generado podremos ver que consta de varias páginas:

Summary of OLS Results - Model Variables

Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]	Robust_SE	Robust_t	Robust_Pr [b]	VIF [c]
Intercept	28,368607	6,573639	4,315511	0,000142*	7,289766	3,891566	0,000473*	--------
Y_WEB	-0,000002	0,000002	-1,608000	0,117657	0,000002	-1,444661	0,158274	1,121760
MDE_EXT	-0,000856	0,000709	-1,206512	0,236464	0,000626	-1,367496	0,180998	1,034137
ORIENTACIO	-0,000731	0,000925	-0,790769	0,434901	0,000912	-0,801888	0,428533	1,118249

En la primera podemos ver el coeficiente para cada variable explicativa. Observando nuestros valores, nuestra ecuación será:

T_med=-0,000002*Latitud-0,000856*Altitud-0,000731*orientacion+28,368607

Si el signo es negativo, como es nuestro caso, la relación es negativa. A mayor cota del MDE, menor temperatura media.

En teoría, a mayor valor, mayor ponderación tiene en la explicación de la temperatura. En nuestro caso, será la altitud, seguida de la orientación la que mayor explicación da a la variable temperatura media (aunque tienen un peso poco definitorio). La explicación de que ninguna de las variables explicativas son definitorias es que en el **estadístico T**, que se utiliza para evaluar si una variable explicativa es estadísticamente significativa o no, <u>la hipótesis nula es que el coeficiente es, en todos sus propósitos, igual a cero (y, como consecuencia, no ayuda al modelo)</u> y en nuestro caso, al no tener el asterisco es que no son definitorias. Por ejemplo, cuando la probabilidad o la probabilidad robusta (valor p) es muy pequeña, la posibilidad de que el coeficiente sea esencialmente cero también es pequeña.

El VIF (Factor de inflación de la varianza) mide la redundancia entre las variables explicativas. Como regla práctica, las variables explicativas asociadas con los valores del VIF mayores que 7,5 aproximadamente deben quitarse (de a uno por vez) del modelo de regresión.

En la siguiente página se nos muestra el diagnóstico o evaluación de los mínimos cuadrados ordinarios:

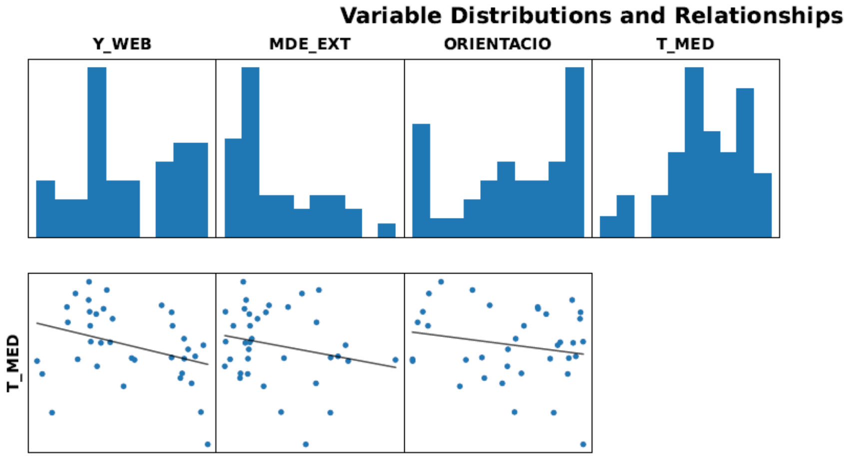
OLS Diagnostics

Input Features	REDAREX_ETSR89h30_clim	Dependent Variable	T_MED
Number of Observations	36	Akaike's Information Criterion (AICc)['d']	73,461295
Multiple R-Squared['d']	0,162768	Adjusted R-Squared['d']	0,084277
Joint F-Statistic['e']	2,073725	Prob(>F), (3,32) degrees of freedom	0,058282
Joint Wald Statistic['e']	6,475334	Prob(>chi-squared), (3) degrees of freedom	0,090640
Koenker (BP) Statistic['f']	1,907698	Prob(>chi-squared), (3) degrees of freedom	0,591784
Jarque-Bera Statistic['g']	1,817409	Prob(>chi-squared), (2) degrees of freedom	0,403046

Tanto el índice **estadístico F conjunto** como el índice **estadístico de Wald conjunto** son medidas de la importancia estadística general del modelo. El índice estadístico F conjunto es confiable únicamente cuando el índice estadístico de Koenker (BP) no es estadísticamente significativo. Si el índice estadístico de Koenker (BP) es significativo, debe consultar el índice estadístico de Wald conjunto para determinar la importancia general del modelo. La hipótesis nula para estas dos pruebas es que las variables explicativas del modelo no son efectivas. Para un nivel de confianza del 95 por ciento, un valor p (probabilidad) menor que 0,05 indica un modelo estadísticamente significativo. En nuestro caso, ninguno de los valores de p es menor de 0.05, por lo tanto el modelo no es significativo.

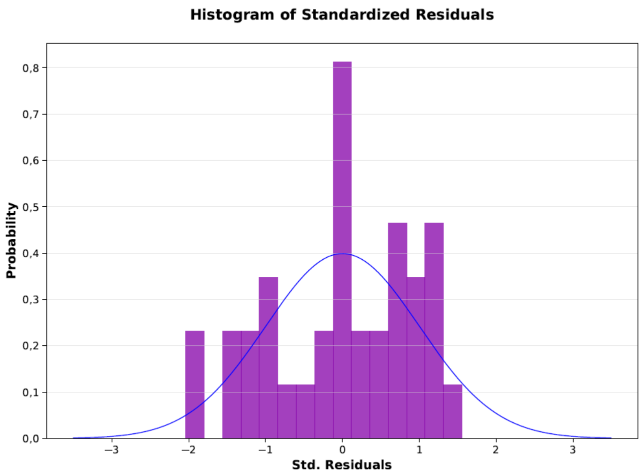
El índice **estadístico de Koenker (BP)** es una prueba para determinar si las variables explicativas del modelo tienen una relación consistente con la variable dependiente, tanto en el espacio geográfico como en el espacio de datos. Cuando el modelo es consistente en el espacio geográfico, los procesos espaciales representados por las variables explicativas se comportan de la misma manera en cualquier parte del área de estudio (los procesos son estacionarios). Cuando el modelo es consistente en el espacio de datos, la variación en la relación entre los valores previstos y cada variable explicativa no cambia cuando cambian las magnitudes de la variable explicativa (no hay heterocedasticidad en el modelo). La hipótesis nula para esta prueba es que el modelo es estacionario. Para un nivel de confianza del 95 por ciento, un valor p (probabilidad) menor que 0,05 indica una heterocedasticidad o no estacionariedad estadísticamente significativa. En nuestro caso, sí existe estacionariedad.

En la siguiente página se nos muestra la correlación entre cada una de las variables explicativas y la variable modelada (dependiente):



En nuestro caso podemos observar como existe una gran dispersión de todas las variables explicativas con respecto a la Temperatura media.

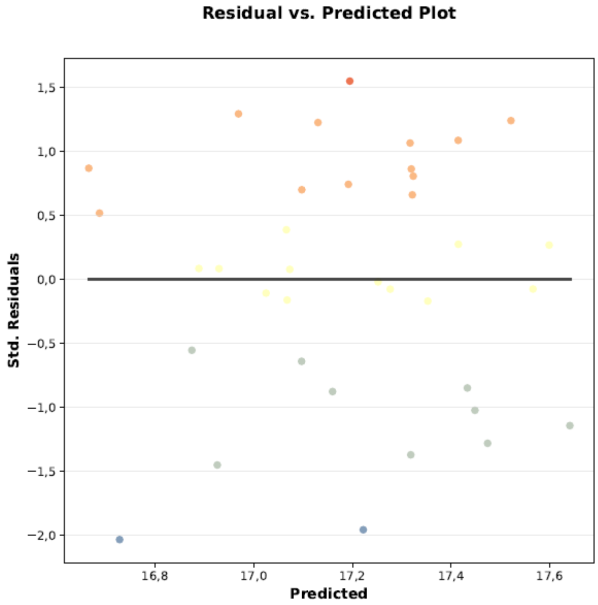
La página siguiente muestra el histograma de los residuales del modelo:



Las barras del histograma muestran la distribución real, y la línea azul superpuesta sobre el histograma muestra la forma que tendría el histograma si los residuales, de hecho, se distribuyeran normalmente.

La forma de este histograma está estrechamente relacionado con el índice **estadístico de Jarque-Bera**, que indica si los residuales (los valores de la variable dependiente observada o conocida menos los valores previstos o estimados) se distribuyen normalmente o no. La hipótesis nula para esta prueba es que los residuales son distribuidos normalmente, por lo tanto, si construyera un histograma de dichos residuales, se parecerían a la curva de Bell o a la distribución gaussiana. Cuando el valor p (probabilidad) de esta prueba es bajo (menor que 0,05 para un nivel de confianza del 95 por ciento, por ejemplo), los residuales no son distribuidos normalmente. Ese es nuestro caso, como se puede ver tanto en la gráfica como en el p-valor del Jarque-Bera.

Finalmente, en la penúltima página se muestra el gráfico la relación entre los residuales del modelo y los valores previstos.

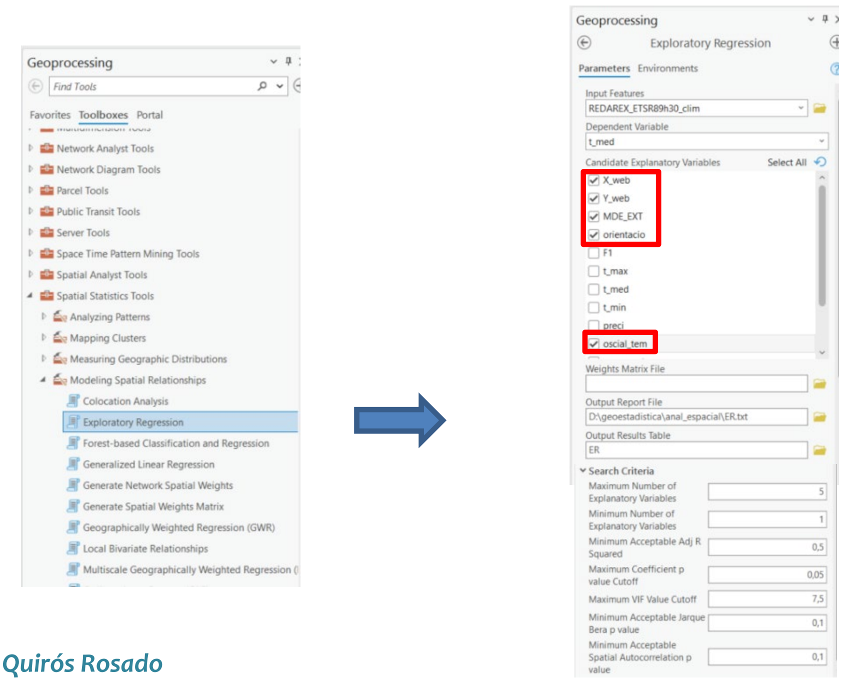


El gráfico indica las predicciones excesivas e insuficientes del modelo en función de la variable en cuestión. Lo correcto es que sea un gráfico disperso. Si hubiera agrupaciones de valores, significaría que el modelo está haciendo buenas predicciones en ubicaciones con altos o bajos valores de temperatura.

Lo normal es construir diferentes modelos combinando distintas variables independientes, y ver cuál es el mejor. Una forma recomendable de hacerlo es utilizar el **Criterio de información de Akaike corregido (AICc)** de la primera página, para comparar diferentes modelos. El modelo con el valor del AICc más pequeño es el mejor (es decir, teniendo en cuenta la complejidad del modelo, el modelo con el AICc más pequeño se ajusta mejor a los datos observados). En nuestro caso el AICc es 73,461295, valor no muy elevado.

4.5 Regresión Exploratoria

Para realizar la regresión exploratoria iremos a la herramienta correspondiente:



En los criterios de búsqueda tenemos que ver que un modelo OLS es adecuado cuando tiene:

- Variables explicativas donde todos los coeficientes son estadísticamente significativos.
- Los coeficientes reflejan la relación esperada, o al menos una justificable, entre cada variable explicativa y la variable dependiente.
- Las variables explicativas no son redundantes; los valores de VIF han de ser menores de 7,5.
- Los valores residuales tienen distribución normal, (el p-valor Jarque-Bera no es estadísticamente significativo).
- Tiene distribución aleatoria de las predicciones excesivas e insuficientes (el p-valor de la autocorrelación espacial no es estadísticamente significativo).

Como resultado nos dará un txt con la siguiente información:

Primero nos dará los mejores valores en función del número de variables explicativas:

```
******************************************************************************
Choose 1 of 5 Summary
   Highest Adjusted R-Squared Results
AdjR2  AICc   JB K(BP)  VIF   SA   Model
 0,09 70,23 0,33  0,51 1,00 0,47  -Y_WEB**
 0,04 72,02 0,43  0,69 1,00 0,00  -OSCIAL_TEM
 0,02 72,88 0,60  0,41 1,00 0,00  -MDE_EXT
       Passing Models
AdjR2 AICc JB K(BP) VIF SA   Model
******************************************************************************
Choose 2 of 5 Summary
           Highest Adjusted R-Squared Results
AdjR2  AICc   JB K(BP)  VIF   SA   Model
 0,48 51,24 0,84  0,27 2,51 0,00  -MDE_EXT***  -OSCIAL_TEM***
 0,16 68,75 0,22  0,52 1,01 0,57  -Y_WEB**  -OSCIAL_TEM*
 0,09 71,45 0,44  0,46 1,01 0,44  -Y_WEB*  -MDE_EXT
       Passing Models
AdjR2 AICc JB K(BP) VIF SA   Model
******************************************************************************
Choose 3 of 5 Summary
                 Highest Adjusted R-Squared Results
AdjR2  AICc   JB K(BP)  VIF   SA   Model
 0,60 43,58 0,78  0,26 2,51 0,05  -Y_WEB***  -MDE_EXT***  -OSCIAL_TEM***
 0,48 53,20 0,88  0,54 2,52 0,00  -X_WEB  -MDE_EXT***  -OSCIAL_TEM***
 0,47 53,66 0,92  0,26 2,65 0,00  -MDE_EXT***  -ORIENTACIO  -OSCIAL_TEM***
       Passing Models
AdjR2 AICc JB K(BP) VIF SA   Model
******************************************************************************
Choose 4 of 5 Summary
                       Highest Adjusted R-Squared Results
AdjR2  AICc   JB K(BP)  VIF   SA   Model
 0,61 44,57 0,57  0,38 2,52 0,62  -X_WEB  -Y_WEB***  -MDE_EXT***  -OSCIAL_TEM***
 0,59 46,16 0,92  0,26 2,70 0,09  -Y_WEB***  -MDE_EXT***  +ORIENTACIO  -OSCIAL_TEM***
 0,47 55,56 0,93  0,60 2,67 0,00  -X_WEB  -MDE_EXT***  -ORIENTACIO  -OSCIAL_TEM***
       Passing Models
AdjR2 AICc JB K(BP) VIF SA   Model
******************************************************************************
Choose 5 of 5 Summary
                           Highest Adjusted R-Squared Results
AdjR2  AICc   JB K(BP)  VIF   SA   Model
 0,60 47,56 0,62  0,37 2,71 0,68  -X_WEB  -Y_WEB***  -MDE_EXT***  +ORIENTACIO  -OSCIAL_TEM***
       Passing Models
AdjR2 AICc JB K(BP) VIF SA   Model
******************************************************************************
```

En nuestro caso, como hemos indicado mínimo 1 variable explicativa y máximo 5, nos da los 5 posibles casos de combinaciones de variables eligiendo o bien 1, 2, 3, 4 o 5. Para cada número de variables nos da un resumen de los tres mejores modelos según R^2 y sus estadísticos correspondientes.

El siguiente apartado indica el resumen global de la regresión exploratoria:

```
*************** Exploratory Regression Global Summary (T_MED) ***************

              Percentage of Search Criteria Passed
Search Criterion                      Cutoff Trials # Passed % Passed
Min Adjusted R-Squared              > 0,50     31        4    12,90
Max Coefficient p-value             < 0,05     31        3     9,68
Max VIF Value                       < 7,50     31       31   100,00
Min Jarque-Bera p-value             > 0,10     31       31   100,00
Min Spatial Autocorrelation p-value > 0,10     13        5    38,46
```

Según los requisitos que indicamos en el cálculo, de los 31 intentos de modelos, sólo todos los modelos cumplen con el VIF y con el Jarque Bera y muy pocos modelos cumplen con el criterio del R2, del p-valor del coeficiente y de la autocorrelación espacial.

La tercera sección sería la del resumen de significancia de cada variable:

```
      Summary of Variable Significance
Variable   % Significant % Negative % Positive
Y_WEB              68,75     100,00       0,00
MDE_EXT            50,00     100,00       0,00
OSCIAL_TEM         50,00     100,00       0,00
X_WEB               0,00     100,00       0,00
ORIENTACIO          0,00      87,50      12,50
```

El listado se ordena con la proporción de veces en la que cada variable explicativa fue estadísticamente significativa. Las primeras variables de la lista son las más significativas. También se puede ver la estabilidad de las variables analizando las columnas % negativo y % positivo. Los indicadores más potentes serán coherentemente estables (principalmente negativos o principalmente positivos).

La siguiente sección se refiere a la multicolinealidad:

```
   Summary of Multicollinearity
Variable     VIF Violations Covariates
X_WEB       1,07          0 --------
Y_WEB       1,14          0 --------
MDE_EXT     2,55          0 --------
ORIENTACIO  1,24          0 --------
OSCIAL_TEM  2,71          0 --------
```

Aquí se indica cuántas veces se incluyó cada variable explicativa en modelos con alta multicolinealidad, y el resto de variables explicativas que también se incluyeron en dichos modelos. Cuando dos (o más) variables explicativas se encuentran juntas con frecuencia en los modelos con alta multicolinealidad, indica que esas variables aportan información redundante. En nuestro caso, no hay ninguna variable redundante. Este hecho lo ratifica tanto el bajo valor del VIF, como el número de violaciones (sin covariables).

Finalmente, se muestran los resúmenes de diagnóstico adicionales.

```
                           Summary of Residual Normality (JB)
      JB    AdjR2      AICc     K(BP)       VIF        SA   Model
0,933899 0,469560 55,558638 0,600564 2,666747 0,000031  -X_WEB  -MDE_EXT***  -ORIENTACIO  -OSCIAL_TEM***
0,918491 0,591478 46,156810 0,259763 2,696077 0,086986  -Y_WEB***  -MDE_EXT***  +ORIENTACIO  -OSCIAL_TEM***
0,915606 0,471683 53,660678 0,264638 2,649854 0,000001  -MDE_EXT***  -ORIENTACIO  -OSCIAL_TEM***

------------------------------------------------------------------------------

                     Summary of Residual Spatial Autocorrelation (SA)
      SA    AdjR2      AICc        JB    K(BP)       VIF   Model
0,683516 0,597348 47,558840 0,615894 0,373671 2,710795  -X_WEB  -Y_WEB***  -MDE_EXT***  +ORIENTACIO  -OSCIAL_TEM***
0,622775 0,609138 44,566007 0,574533 0,379686 2,518871  -X_WEB  -Y_WEB***  -MDE_EXT***  -OSCIAL_TEM***
0,574421 0,160164 68,745149 0,215384 0,520361 1,014479  -Y_WEB**  -OSCIAL_TEM*
```

Las dos secciones muestran los p-valores de Jarque-Bera más elevados (Resumen de normalidad residual) y los p-valores de I de Moran global más elevados (Resumen de autocorrelación residual) de los tres mejores modelos.

Estos resúmenes son útiles cuando no se tiene ningún modelo que pase la prueba, como es nuestro caso, y se desea ver lo lejos que está de haber distribuido normalmente los residuales o residuales que estén libres de la autocorrelación espacial estadísticamente significativa. Por ejemplo, en nuestro caso, el problema es que pocos modelos pasan el criterio de la autocorrelación espacial (I-Moran), que hemos dicho que ha de ser mayor de 0,1. Vemos que al menos hay un modelo (marcado en rojo) que estaría casi a punto de cumplirlo.

```
                              Summary of Residual Normality (JB)
     JB      AdjR2      AICc     K(BP)       VIF       SA   Model
0,933899 0,469560 55,558638 0,600564 2,666747 0,000031  -X_WEB  -MDE_EXT***  -ORIENTACIO  -OSCIAL_TEM***
0,918491 0,591478 46,156810 0,259763 2,696077 0,086986  -Y_WEB***  -MDE_EXT***  +ORIENTACIO  -OSCIAL_TEM***
0,915606 0,471683 53,660678 0,264638 2,649854 0,000001  -MDE_EXT***  -ORIENTACIO  -OSCIAL_TEM***

------------------------------------------------------------------------------

                           Summary of Residual Spatial Autocorrelation (SA)
     SA      AdjR2      AICc      JB     K(BP)       VIF    Model
0,683516 0,597348 47,558840 0,615894 0,373671 2,710795  -X_WEB  -Y_WEB***  -MDE_EXT***  +ORIENTACIO  -OSCIAL_TEM***
0,622775 0,609138 44,566007 0,574533 0,379686 2,518871  -X_WEB  -Y_WEB***  -MDE_EXT***  -OSCIAL_TEM***
0,574421 0,160164 68,745149 0,215384 0,520361 1,014479  -Y_WEB**  -OSCIAL_TEM*
```

Así sabemos que la combinación de variables (-)Latitud, (-)Elevación, (+)Orientación y (-)Oscilación de temperatura, formaría el modelo que estarían a punto de cumplir con los requisitos especificados.

Por otro lado, en la tabla de resultados de salida que hemos guardado podremos ver un listado con los modelos (en nuestro caso tres) que cumplen el valor límite máximo de p-valor de coeficiente y Valor límite de valor VIF.

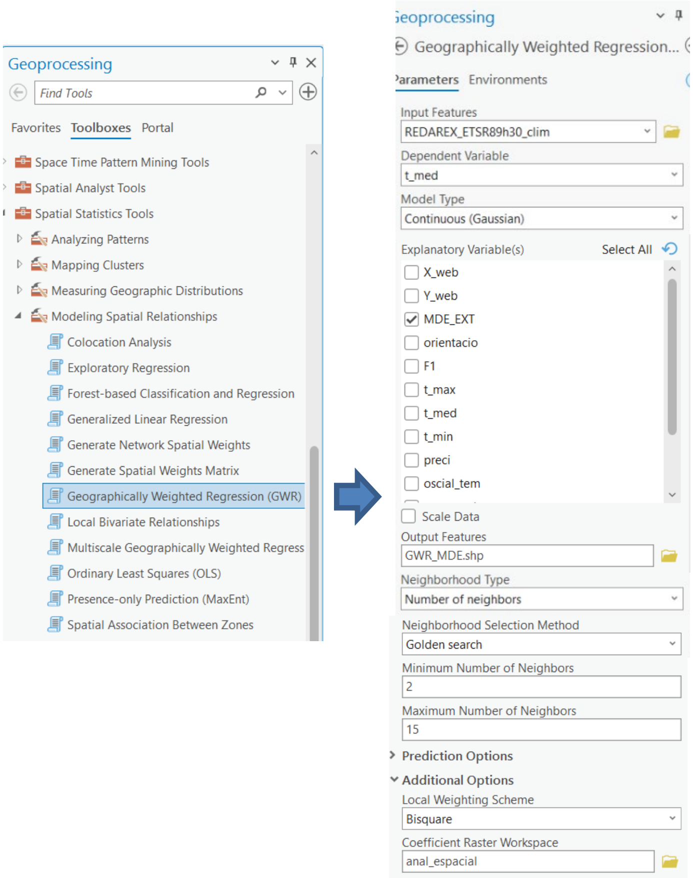
OID	Field1	RunID	AdjR2	AICc	JB	K_BP	MaxVIF	SA	NumVars	X1	X2	X3
0	0	1	0,088379	70,232173	0,327955	0,505322	1	0,466358	1	Y_WEB		
1	0	2	0,483618	51,2362	0,842414	0,273048	2,505512	0	2	MDE_EXT	OSCIAL_TEM	
2	0	3	0,600691	43,582068	0,782727	0,261379	2,511411	0,047323	3	Y_WEB	MDE_EXT	OSCIAL_TEM

En nuestro caso, a pesar de que, según el informe, ningún modelo cumplía los requisitos, tenemos tres modelos. Lo más adecuado es ordenar los modelos por sus valores AICc. Mientras más bajo sea el valor de AICc, mejor es el modelo a realizar.

OID	Field1	RunID	AdjR2	AICc	JB	K_BP	MaxVIF	SA	NumVars	X1	X2	X3
2	0	3	0,600691	43,582068	0,782727	0,261379	2,511411	0,047323	3	Y_WEB	MDE_EXT	OSCIAL_TEM
1	0	2	0,483618	51,2362	0,842414	0,273048	2,505512	0	2	MDE_EXT	OSCIAL_TEM	
0	0	1	0,088379	70,232173	0,327955	0,505322	1	0,466358	1	Y_WEB		

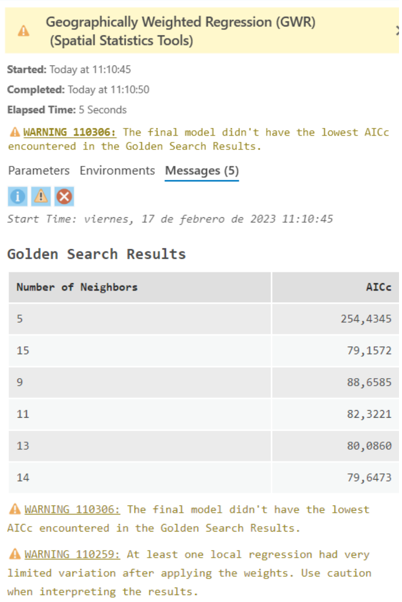
Por tanto, el mejor modelo sería el de tres variables teniendo en cuenta la latitud, la elevación y la oscilación de la temperatura.

4.6 Regresión ponderada geográficamente (GWR)



Todos los parámetros con los que hemos configurado el cálculo los hemos visto en teoría salvo el parámetro Método de selección de vecindad, que especifica cómo se determina el tamaño de la vecindad (la distancia o la cantidad de vecinos utilizados en realidad). La vecindad seleccionada con la opción Búsqueda dorada (*Golden search*) se basa en minimizar el valor del criterio de información de Akaike (AICc).

Nos saldrá, como era de esperar, que no hay modelos que superen el AIC,



En la sección del modelo de diagnóstico nos aparecerá la siguiente información:

- **R^2**: Medida de la bondad de ajuste. Se puede interpretar como la proporción de varianza de la variable dependiente que da cuenta el modelo de regresión.
- **AdjR^2**: Es un R2 ajustado para compensar el número de variables en un modelo. Sin embargo, al realizar este ajuste, pierde la interpretación del valor como una proporción de la varianza explicada y es preferible tomar el AIC como medio de comparación de modelos.
- **AIC**: No es una medida absoluta de la bondad de ajuste, pero es útil para comparar los modelos con distintas variables explicativas. El modelo con el valor AIC más bajo proporciona un mejor ajuste para los datos observados. En nuestro caso vemos como no ha escogido el modelo con mayor AIC.
- **σ^2** (Sigma cuadrado): es la estimación de mínimos cuadrados de la varianza (desviación estándar cuadrada) para los residuales. Se prefieren los valores más pequeños de esta estadística.
- **σ^2MLE**: Es la estimación de máxima probabilidad (MLE) de la varianza de los residuales. Se prefieren los valores más pequeños de esta estadística.
- **Grados de libertad efectivos**: este valor refleja un equilibrio entre la varianza de los valores ajustados y la influencia en las estimaciones de coeficiente, y se relaciona con la opción de tamaño de vecindad. El alto valor obtenido en nuestro caso, indica que tenemos un tamaño de vecindad muy grande; las estimaciones de coeficiente local tendrán una pequeña varianza, pero estarán un poco influenciadas.

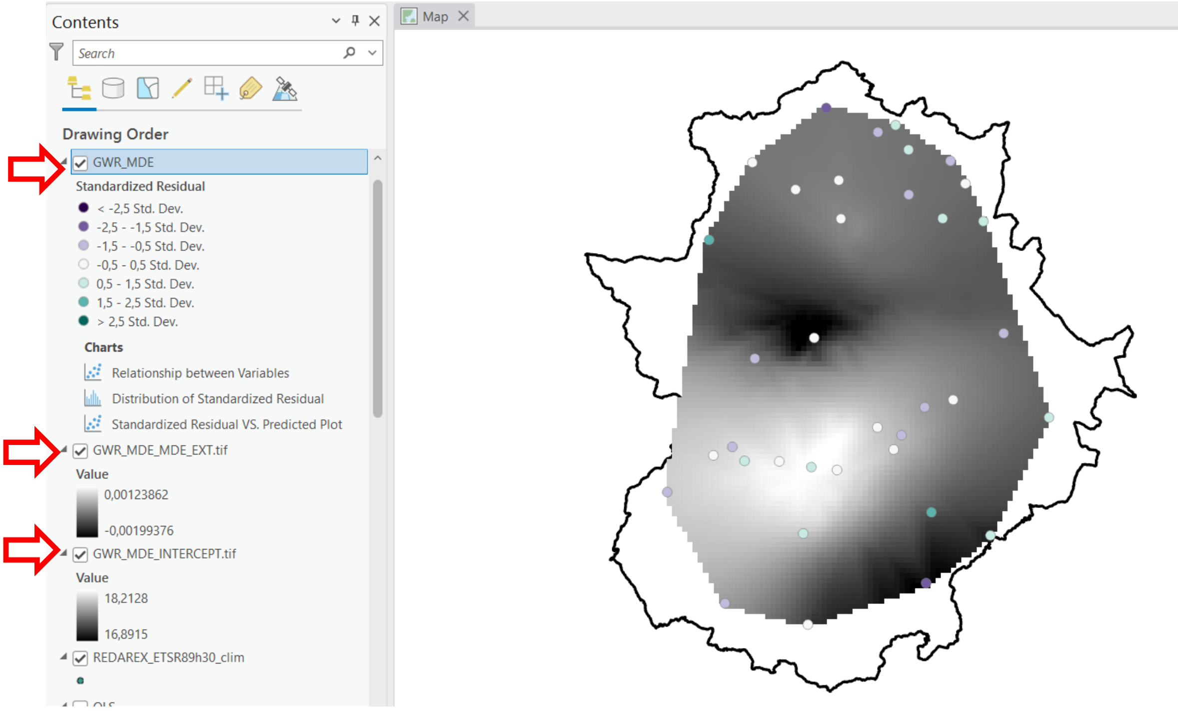
Analysis Details

Number of Features	36
Dependent Variable	T_MED
Explanatory Variables	MDE_EXT
Number of Neighbors	14

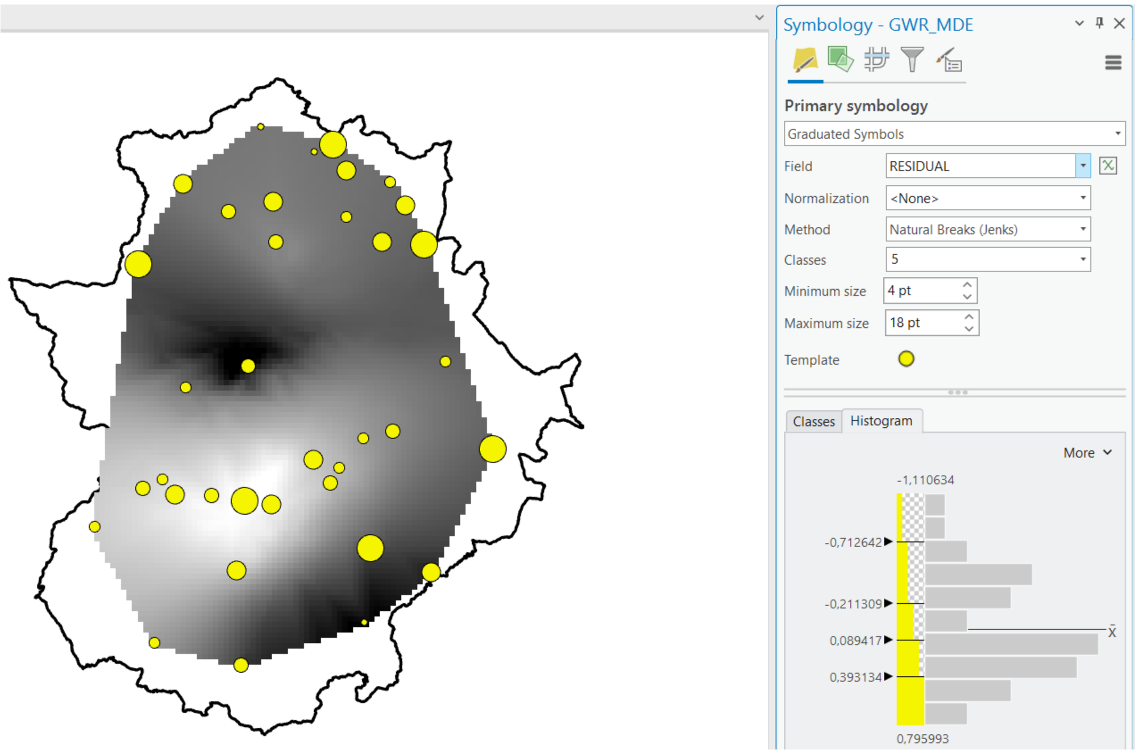
Model Diagnostics

R2	0,4871
AdjR2	0,1832
AICc	79,6473
Sigma-Squared	0,3099
Sigma-Squared MLE	0,1978
Effective Degrees of Freedom	22,9786
Adjusted Critical Value of Pseudo-t Statistics	2,7483

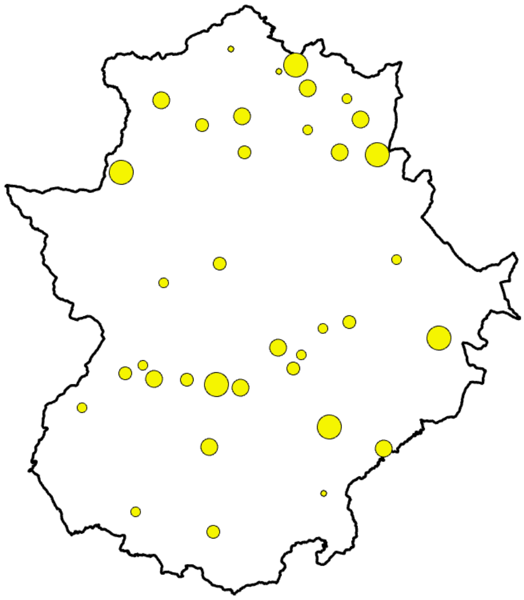
Además, me habrá generado tres capas más de información.



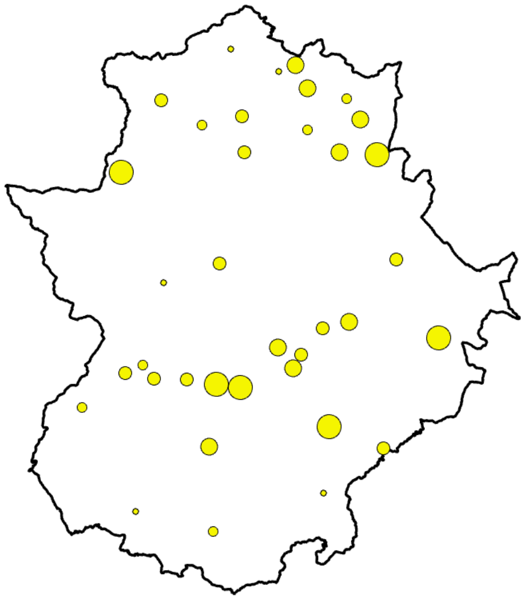
Si nos centramos en la capa vectorial, podremos, igual que en el cálculo de OLS, representar cada estación en función de los residuales.



Podríamos así comparar los residuales de esta técnica geoestadística con los del OLS y ver si los mayores errores se producen en las mismas estaciones.



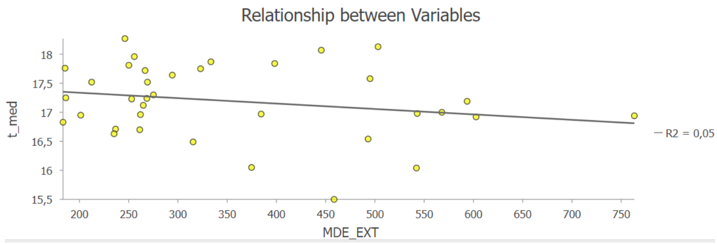
GWR

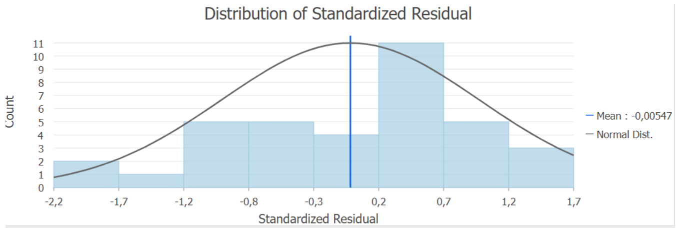


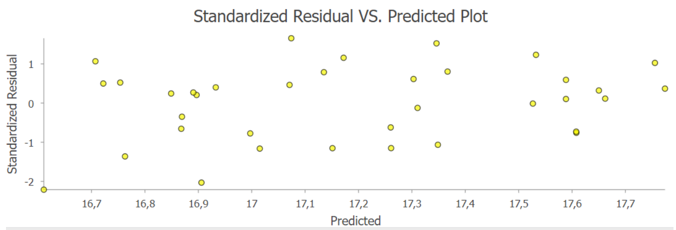
OLS

Comprobamos que los residuales salen muy parecidos.

Además, se nos han generado varios gráficos En ellos podemos ver que el ajuste no es muy elevado:

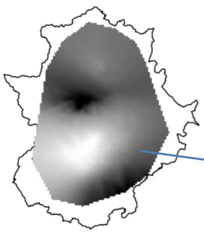




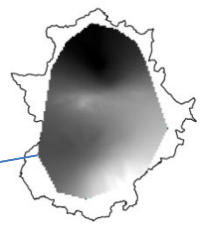


Finalmente, analizaremos los ráster que nos ha generado la herramienta:

Por un lado, el ráster de intercepción, y por otro, el de la variable explicativa. De tal manera que la combinación de ambos nos daría el valor de la temperatura estimada en cada zona:



$$Tm = \beta 1 \cdot elevación + \epsilon$$



Ejercicio propuesto:

Con cualquier otra variable dependiente e independientes o explicativas, de las medidas en las estaciones de Redarex, hacer los cálculos de OLS y GWR.

3. Bibliografía

Brunsdon, C., Fotheringham, A. S., y Charlton, M. E. (1996). "Geographically weighted regression: a method for exploring spatial nonstationarity". Geographical analysis, 28(4), 281-298.

Chirivella González, V. (2015). Hipótesis en el modelo de regresión lineal por Mínimos Cuadrados Ordinarios.

Fotheringham, Stewart A., Chris Brunsdon y Martin Charlton. Geographically Weighted Regression: the analysis of spatially varying relationships. John Wiley & Sons, 2002.

Giraldo, R. (2011). Introducción a la geoestadística. Universidad Nacional de Colombia.

Gollini, I., Lu, B., Charlton, M., Brunsdon, C., y Harris, P. (2013). GWmodel: an R package for exploring spatial heterogeneity using geographically weighted models. arXiv preprint arXiv:1306.0413.

Ma, Z., Xue, Y., & Hu, G. (2021). Geographically weighted regression analysis for spatial economics data: A Bayesian recourse. International Regional Science Review, 44(5), 582-604.

Mitchell, A. (2005). ESRI Manual for GIS analysis. Volume, 2, 190.

Nakaya, T., Fotheringham, A. S., Brunsdon, C., y Charlton, M. (2005). "Geographically weighted Poisson regression for disease association mapping". Statistics in medicine, 24(17), 2695-2717.

Páez, A., Farber, S., y Wheeler, D. (2011). "A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships". Environment and Planning A, 43(12), 2992-3010.