

Introducción a la Estadística Espacial

Dr. Andrés Farall

April 29, 2022

Motivación

Podemos mencionar al menos dos características fundamentales para considerar a la estadística espacial como un campo de estudio específico.

- El espacio es un factor con una estructura particular relevante para explicar/predecir muchos fenómenos de interés. Tecnicamente, si queremos analizar la variable Y que posee un localización espacial S , generalmente sucede que su comportamiento (distribución D) depende fuertemente del componente espacial S , es decir $D(Y/S) \neq D(Y)$.
- Porque, incluso teniendo en cuenta los factores relevantes X para analizar el fenómeno Y , otra información relevante proviene del comportamiento del fenómeno en un entorno cercano. Tecnicamente, para la variable Y_1 con localización espacial S_1 , generalmente sucede que su comportamiento depende de Y_2 con localización espacial S_2 , es decir $D(Y_1/X, S, Y_2) \neq D(Y_1/X, S)$, si S_1 está cerca de S_2 .

Procesos Estocásticos Espaciales

De manera general podemos definir a un proceso estocástico espacial como una colección de variables o vectores aleatorios Z indexados en alguna región S del espacio R^d (normalmente con $d=2$ o $d = 3$), es decir

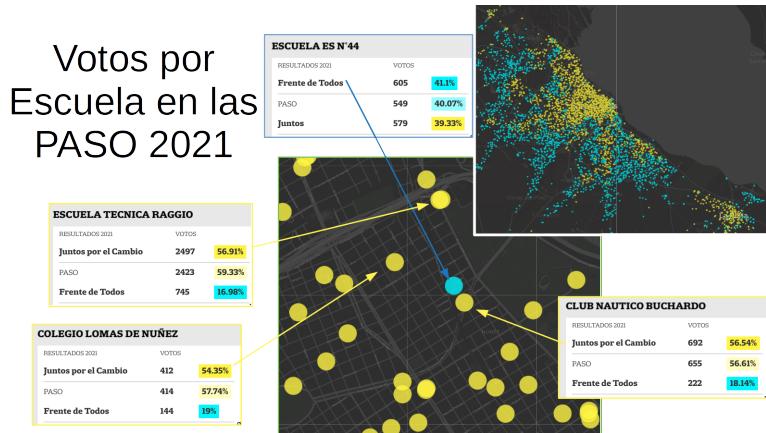
$$\{Z(s) : s \in S \subseteq R^d\}$$

En estos procesos hay dos fuentes bien distintas (y potencialmente relacionadas) de variación o aleatoriedad. La primera es la localización espacial de los casos (o eventos), denotada por s . La segunda es la valuación de la variable o vector aleatorio Z .

Para poder caracterizar completamente el proceso, necesitamos conocer la distribución conjunta de cualquier conjunto finito de estas variables o vectores $Z(s_1), Z(s_2) \dots Z(s_n)$, es decir conocer

$$P(Z(s_1) \in A_1, Z(s_2) \in A_2 \dots Z(s_n) \in A_n)$$

Un ejemplo concreto de una realización de este tipo de procesos es el resultado de una votación a nivel de establecimiento, para el cual tenemos la localización de cada establecimiento. La próxima figura muestra los sitios de votación alrededor del conurbano bonaerense (arriba a la derecha), junto con un detalle espacial del noreste del barrio de Nuñez (centro de la imagen). Asociado a cada punto tenemos la cantidad de votos del sitio y las proporciones resultantes del sufragio. Llama poderosamente la atención el resultado del punto celeste, que no respeta la “tendencia espacial” de las proporciones de votos de la zona.



En este caso particular, las localizaciones puntuales $s \in S$ pueden ser pensadas como un proceso puntual aleatorio, en tanto que a priori podríamos no saber cuáles de todos los establecimientos potencialmente utilizables serán efectivamente utilizados para el comicio. En cuanto a la variable aleatoria Z podemos pensar en la proporción de votos de un determinado partido político. O, podemos pensar en un vector aleatorio Z que contemple todas las proporciones de votos de todos los partidos. Algunas preguntas que surgen naturalmente son:

- ¿ Se hallan las localizaciones s concentradas espacialmente ?
- ¿ Como se comporta la variable Z ?
- ¿ Existe una relación entre las localizaciones s y los valores de la variable Z ?

Nótese que la primer pregunta involucra sólo a las localizaciones. La segunda pregunta comprende exclusivamente a la variable Z . En tanto que la tercer pregunta relaciona a las localizaciones con la variable Z .

La situación más sencilla de proceso estocástico espacial es aquella en la cual sólo tenemos información (o sólo tenemos interés) sobre la localización espacial de los eventos (primer pregunta). Esto puede ser pensado como un caso particular del proceso general mencionado anteriormente, pues alcanza con definir $Z(s) = 1$ si el evento ocurrió en la posición s y $Z(s) = 0$ si no hay evento en la posición s . Pese a esto, a este caso tan particular se lo conoce

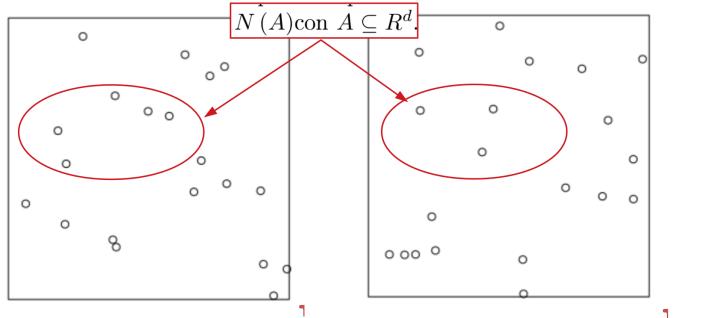
como Procesos Espaciales Puntuales (Point Pattern), y se los trabaja de manera totalmente diferente, como veremos a continuación.

Procesos Espaciales Puntuales

Un proceso estocástico puntual \mathbf{X} es un proceso que genera conjuntos finitos aleatorios de puntos en R^d . Una definición más formal de un proceso estocástico puntual \mathbf{X} proviene de pensar que para toda región compacta $A \subseteq R^d$ puede definirse una función $N(A) = \# \{X \cap A\}$, que cuenta la cantidad de eventos en la región, y que se comporta como una variable aleatoria con una cierta distribución. De esta forma, el proceso queda caracterizado por las distribuciones de las variables aleatorias $N(A)$ con $A \subseteq R^d$.

Sorprendentemente, la distribución del proceso \mathbf{X} sólo depende de las probabilidades de eventos dicotómicos de presencia/ausencia, es decir, que alcanza con conocer el comportamiento de $V(A) = P(N(A) = 0)$ para toda región $A \subseteq R^d$.

La siguiente figura muestra dos realizaciones de un proceso puntual sobre una ventana $S \subset R^d$.



Es fundamental entender que realizaciones del mismo proceso pueden ser (y en general lo son) muy distintas !

Proceso Binomial

El proceso más sencillo que puede planterse es el proceso binomial.

El proceso binomial es aquel que genera n eventos de manera independiente sobre una región compacta $S \subseteq R^d$, donde cada uno de los eventos obedecen a la distribución espacial (pdf) $f(x)$ con $x \in S$. De esta forma, la variable aleatoria $N(A)$ tiene una distribución binomial $b(n, p)$ con la probabilidad $p = \int_A f(x) dx$.

El caso particular más simple de proceso binomial consiste un fijar una distribución espacial constante para los eventos $f(x) = c$ con $x \in S$.

¿ Cómo generamos una muestra de este proceso ?

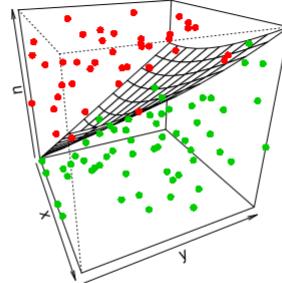
Muy facil. Si la región S es rectangular $S = [a, b] \times [c, d]$, basta con generar n pares de observaciones uniformes en los intervalos $[a, b]$ y $[c, d]$ de manera independiente.

¿ Cómo generamos una muestra del proceso binomial para cualquier pdf ?

También es facil. Se basa en el **Principio Fundamental de la Simulación** (no, no es un curso para políticos):

Si $(\mathbf{X}, U) \sim Unif((\mathbf{x}, u) : 0 \leq u \leq f(\mathbf{x}))$ entonces $\mathbf{X} \sim f$

La intuición de este principio puede obtenerse rápidamente del próximo gráfico



En la figura se muestran $N = 100$ eventos generados de manera independiente en el cubo $[0, 1]^3$. Si quisieramos generar eventos en el plano (x, y) que sigan una densidad $f(x)$ que crezca cuadráticamente con ambas variables, representada por la superficie de la figura, alcanza con tonar los puntos (color verde) que quedan por debajo de la superficie.

El mismo gráfico sugiere la mecánica de simulación: la técnica de aceptación-rechazo por Monte Carlo. Genramos una a una observaciones $\sim Unif(\mathbf{x}, u)$ en el cubo, y nos quedamos con (aceptamos) las observaciones que satisfacen $0 \leq$

$u \leq f(\mathbf{x})$, y rechazamos el resto. El proceso se interrumpe cuando alcanzamos n observaciones aceptadas.

El mecanismo antes descripto produce realizaciones de un proceso binomial general (bajo cualquier densidad). Nótese que la generación de observaciones en un soporte más irregular que un rectángulo se obtiene forzando a la función f a tomar valor 0 en los puntos que no están en el soporte. Un caso muy especial es la generación de casos con densidad uniforme en soportes irregulares, en los que se generan casos uniformes en el rectángulo que lo comprende, para luego rechazar las que quedan fuera del soporte.

La intensidad de un proceso

En general, la intensidad de un proceso es una función que, dada una región particular del soporte, nos permite calcular la cantidad esperada de eventos en esa región. En particular para el proceso binomial, esa función es la densidad f del proceso multiplicada por la cantidad de eventos n . Pues la cantidad esperada de eventos en $A \subseteq S \subseteq R^d$ es claramente $E(N(A)) = np_A = n \int_A f(x) dx$.

Proceso Poisson

Pensemos ahora en un proceso similar al binomial, pero reemplazando la función de densidad por una función de intensidad λ , para el cual le pedimos que para toda región compacta $A \subseteq S \subseteq R^d$ cumpla que:

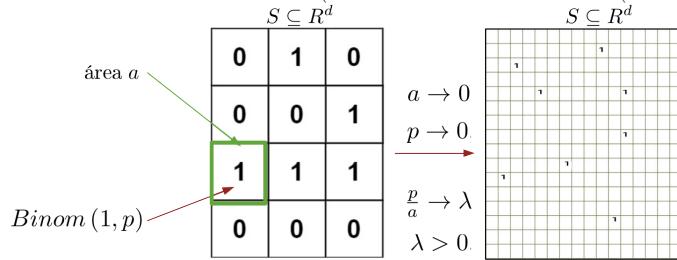
- $N(A) \sim P(\lambda = \int_A \lambda(x) dx)$. O sea que la cantidad de eventos debe comportarse como una poisson.
- Dado $N(A) = n$, $N(B) \sim \text{Binom}(n, \sim \int_B \lambda(x) dx)$ cuando $B \subseteq A$. Es decir que si condicionamos a una cantidad de eventos fija en la ventana que incluye a la región, la cantidad de eventos en la región sigue siendo binomial.
- Dadas las variables aleatorias $N(A)$ y $N(A^*)$, ambas son independientes si $A \cap A^* = \emptyset$, es decir que son regiones disjuntas.

A un proceso de poisson con intensidad constante, cuya función de intensidad $\lambda(x) = c$ no depende del espacio, se lo llama **homogéneo**.

¿ Cómo se llega a un proceso de Poisson con intensidad constante λ ?

Al proceso de poisson puede llegarse de múltiples maneras, una modalidad particularmente intuitiva (al menos para mí), es partir de un proceso discreto en un soporte acotado $S \subseteq R^d$. Armemos una grilla regular con celdas de área a . En cada celda aleatorizamos la ocurrencia del evento mediante una distribución Bernoulli ($\text{Binom}(1, p)$) con probabilidad $p > 0$. Ahora hagamos más fina la grilla, llevando el área de cada celda a cero ($a \rightarrow 0$) y al mismo tiempo haciendo que la probabilidad de ocurrencia del evento también tienda a cero ($p \rightarrow 0$),

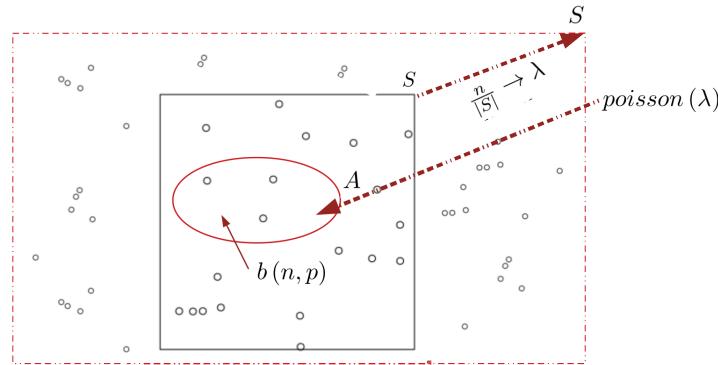
pero de tal forma que la relación entre ambos se mantenga constante y alejada del cero, es decir $\frac{p}{a} \rightarrow \lambda$ con $\lambda > 0$.



En el límite de este ejercicio obtendremos un proceso de Poisson homogéneo con intensidad constante λ .

¿ Cómo se llega a un proceso de Poisson desde el Proceso Binomial ?

Partimos del proceso binomial definido sobre una región compacta $S \subseteq R^d$, donde se generan n eventos de manera uniforme e independiente. Tomando la región $A \subseteq S$, sabemos que la variable aleatoria $N(A)$ tiene una distribución binomial $b(n, p)$ con la probabilidad $p = \frac{|A|}{|S|}$. Agrandemos ahora (al infinito y más allá) la región S y la cantidad de eventos generados, pero de modo tal que $\frac{n}{|S|} \rightarrow \lambda$ constante (la intensidad). La variable aleatoria $N(A)$ tiene ahora una distribución *poisson* (λ) .



¿ Cómo generamos una muestra del proceso poisson para cualquier función de intensidad λ ?

El mecanismo consta de dos pasos:

- Definimos una región compacta $S \subseteq R^d$, y generamos $N \sim P \left(\lambda = \int_S \lambda(x) dx \right)$ como la cantidad de eventos en la región.

- Simulamos n (realización de la v.a. N) eventos según un proceso binomial con densidad $f(x) = \frac{\lambda(x)}{\int_S \lambda(x) dx}$

Importante: El proceso de poisson queda caracterizado por la función de intensidad $\lambda(x)$, la cual NO tiene porque ser una densidad. O sea, NO integra 1. Sin embargo, ambas nociones representan lo mismo, salvo una constante de proporcionalidad, pues $f(x) | S | = \lambda(x)$.

A modo de ejemplo mostramos en el próximo gráfico dos realizaciones de un proceso de poisson inhomogéneo. La primera con función de intensidad:

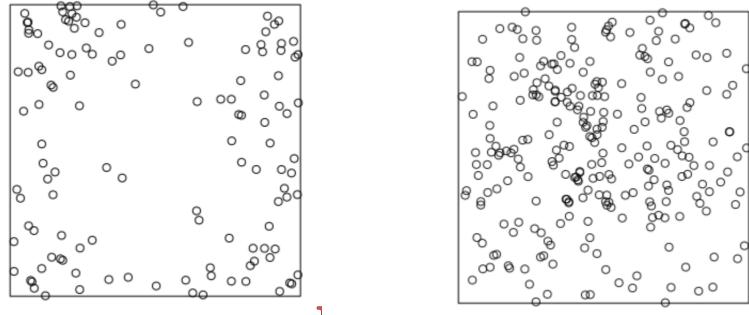
$$\lambda(x,y) = 800 \left(x - \frac{1}{2} \right)^2 + 800 \left(y - \frac{1}{2} \right)^2$$

en la que la intensidad se minimiza en el centro de la ventana $[0,1] \times [0,1]$
La segunda con función de intensidad:

$$\lambda(x,y) = 200 - 800 \left(x - \frac{1}{2} \right)^2 + 200 - 800 \left(y - \frac{1}{2} \right)^2$$

en la que la intensidad se maximiza en el centro de la ventana.

Procesos de Poisson InHomogeneos



Estacionariedad e Isotropía

Presentamos dos conceptos fundamentales de los procesos espaciales.

Estacionariedad: esta propiedad implica que podemos desplazarnos en el soporte del proceso, y el mismo se comporta de la misma manera, es decir

$$\mathbf{X} \sim \mathbf{X} + c \text{ con } c \in R^d$$

Isotropía: esta propiedad implica también invarianza del proceso, pero en este caso en términos de su dirección, o sea

$$\mathbf{X} \sim T\mathbf{X} \text{ con } T \in R^{d \times d} \text{ una matriz de rotación}$$

En un proceso de poisson homogeneo y en un proceso binomial con densidad constante, se cumplen ambas propiedades (salvo por efecto borde, cuidado !).

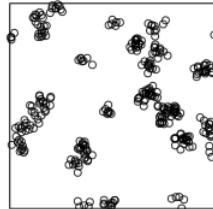
Procesos de Thomas

Los procesos de Thomas son procesos de clusterización. A diferencia de los procesos de poisson y binomial, este nuevo proceso involucra **interacción** (dependencia) entre los eventos. Específicamente, el proceso establece una dependencia positiva entre los casos, haciendo que la localización de un evento en una zona aumente la probabilidad de localización de otros eventos en dicha zona. Este fenómeno tiende a generar clusters de forma natural. Los pasos para simular realizaciones de este proceso son:

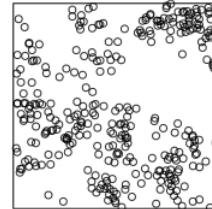
- Generar eventos “padres” provenientes de un proceso de poisson homogéneo con función de intensidad λ .
- Para cada evento “padre” $x \in R^d$ generado en el paso anterior, generar una cantidad poisson de “hijos” con cantidad esperada μ . La localización del “hijo” i -ésimo será $x + \epsilon_i$ con $\epsilon_i \sim N(\mathbf{0}, \sigma^2)$, siendo los ϵ_i independientes entre sí.
- Eliminar los eventos “padre”.

Los parámetros que definen este proceso son: la tasa de intensidad λ , la cantidad esperada de hijos μ y la dispersión σ . La siguiente figura muestra 4 realizaciones de un proceso de Thomas en las que se varían los parámetros μ y σ .

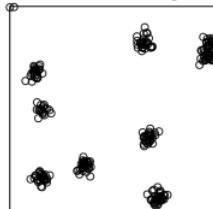
Lambda= 30 Mu= 10 Sigma= 0.02



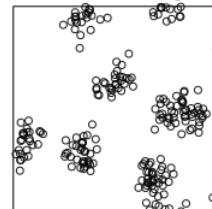
Lambda= 30 Mu= 10 Sigma= 0.05



Lambda= 10 Mu= 30 Sigma= 0.02



Lambda= 10 Mu= 30 Sigma= 0.05

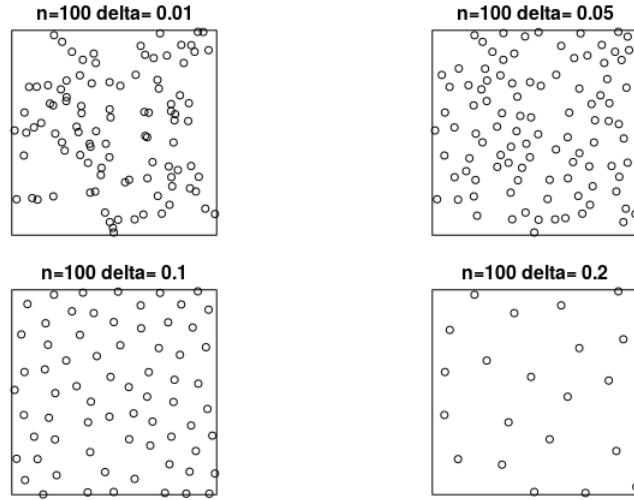


Proceso Secuencial de Inhibición Simple

Este proceso genera eventos con patrones de distribución regular. El proceso para generar n puntos se resume en los siguientes pasos:

1. Defino una ventana $S \subseteq \mathbf{R}^d$ donde tomará lugar el proceso puntual y una distancia δ de inhibición entre puntos.
2. Fijo $k = 1$ y genero un primer punto $X_k = X_1 \in S$ aleatoriamente con densidad uniforme
3. Genero un nuevo punto X_{k+1} con densidad uniforme en $S_{k,\delta} = \{s \in S : \|s - s_i\| > \delta, i = 1 \dots k\}$
4. Si $k + 1 < n$ vuelvo a 3, caso contrario termino

El resultado de un proceso como este puede apreciarse en la siguiente figura.



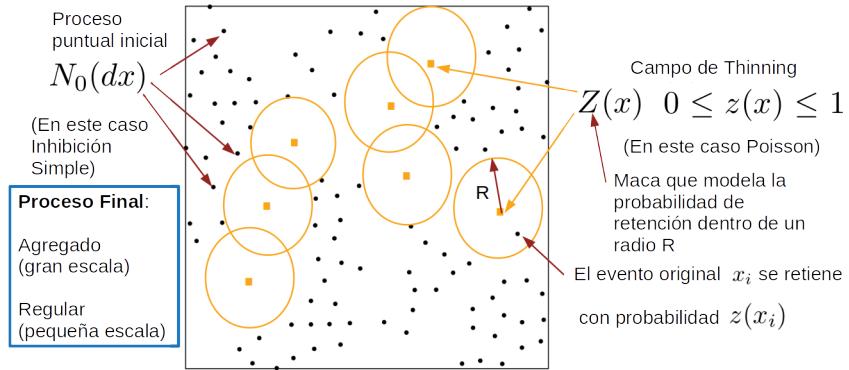
Puede observarse como al aumentar la distancia de inhibición se incrementa la regularidad en el patrón de localización de los eventos.

Proceso de Thinning

El proceso de Thinning simula algunos procesos naturales en los que primero se crean eventos, para luego ser reducidos mediante un proceso de eliminación. El proceso se puede resumir genericamente en los siguientes pasos:

1. Generar una realización de un proceso inicial $N_0(dx)$.
2. Generar una realización de un segundo proceso $Z(X)$ que define que eventos del proceso inicial serán retenidos, y cuales serán eliminados.

Este proceso tiene la potencialidad de generar eventos con patrones de distribución regular a pequeña escala, y con patrones de agregación a gran escala, como se representa en la próxima figura, mediante un caso particular de proceso de Thinning, en el cual el proceso inicial es un proceso de Inhibición Simple, en tanto que el segundo proceso corresponde a un proceso de Poisson homogéneo.

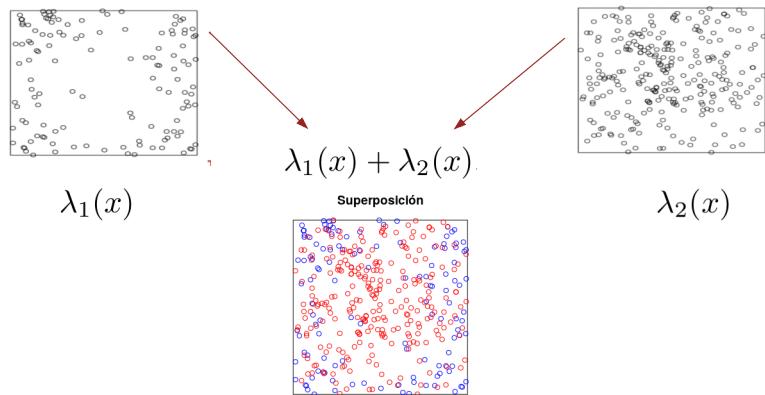


En este caso particular, los eventos del proceso de poisson definen un área de radio R en la cual se retendrán los eventos del proceso inicial con probabilidad $Z(x) = 0$. En casos más generales, la marca de probabilidad de retención puede ser no constante y función del espacio x .

Procesos de Superposición

Como se imaginarán, estos procesos refieren a los procesos resultantes de “superponer” varios procesos en forma simultanea. Un caso particular interesante es la superposición de procesos de Poisson Inhomogeneos, que resulta en otro proceso de Poisson cuya intensidad es la suma de las intensidades de los procesos intervinientes.

La figura siguiente muestra la superposición de las dos realizaciones de los procesos de Poisson inhomogeneos presentados previamente.



Complete Spatial Randomness (CSR)

De forma intuitiva podemos decir que un proceso que satisface la propiedad de Complete Spatial Randomness (CSR) es un proceso que genera eventos **aleatorios independientes** y de manera **uniforme** sobre un area específica.

Estos procesos involucran dos características fundamentales:

- Homogeneidad: todas las subregiones tienen la misma probabilidad de ocurrencia.
- No Interacción: la presencia de un evento en una subregión no condiciona la probabilidad de ocurrencia de otro evento.

El proceso de Poisson homogeneo es EL proceso que respeta la CSR. Por otro lado, un proceso de Poisson inhomogeneo falla en la primer característica, mientras que el proceso de Thomas incumple la segunda característica.

Es improbable que un fenómeno espacial real cumpla en forma exacta la CSR. La relevancia de CSR proviene de servir como una referencia contra la cual se comparan los procesos puntuales empíricos de interés. A continuación veremos algunas métricas que usualmente se aplican para medir la distanica de un proceso real a uno del tipo CSR.

La Función G

La función $G(r)$ computa la proporción de eventos para los cuales su vecino más cercano dista menos que r . Dicho de otra forma, es la función de distribución de la variable (aleatoria) “distancia al vecino más cercano”. Si denominamos d_{ij} a la distancia entre el punto i y el punto j , y $D_i = \min_j \{d_{ij}, \forall j \neq i\}$ la distancia al vecino más cercano del evento i , la función se puede definir así:

$$G(r) = \frac{\#\{D_i : D_i \leq r\}}{n}$$

donde n es la cantidad total de eventos o puntos.

Para un proceso CSR la función G puede calcularse analíticamente, como

$$G(r) = 1 - e^{-\lambda\pi r^2}$$

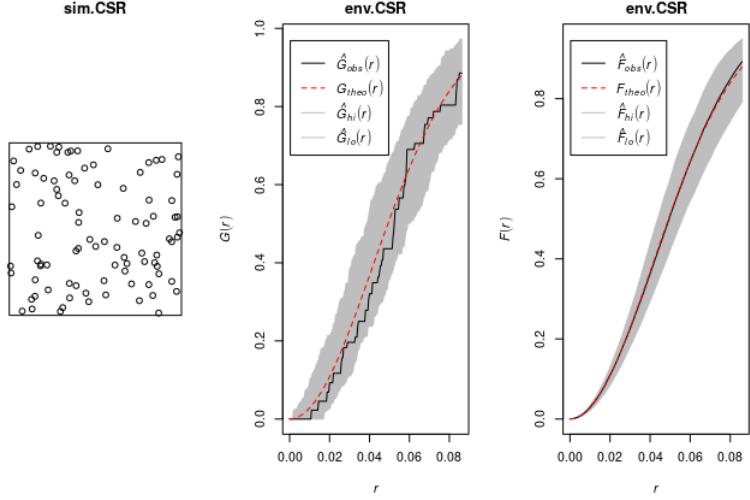
siendo λ la intensidad (constante, por ser SCR) del proceso.

La Función F

Una limitación obvia de la función G es que las distancias a los vecinos más cercanos son medidas sólo desde los mismos eventos. Esto produce por un lado una discretización natural de la curva, ya que los eventos son pocos (en este caso) y no se distribuyen por toda la región. Por otro lado, las distancias a eventos desde zonas vacías no es tenida en cuenta. Una solución (o alternativa) a esto es calcular las distancias a los eventos más cercanos, pero ahora desde cualquier punto arbitrario de la región de interés. Así, la función $F(r)$ computa la función de distribución de la variable (aleatoria) “distancia al evento más cercano” de todos los puntos de la región de análisis.

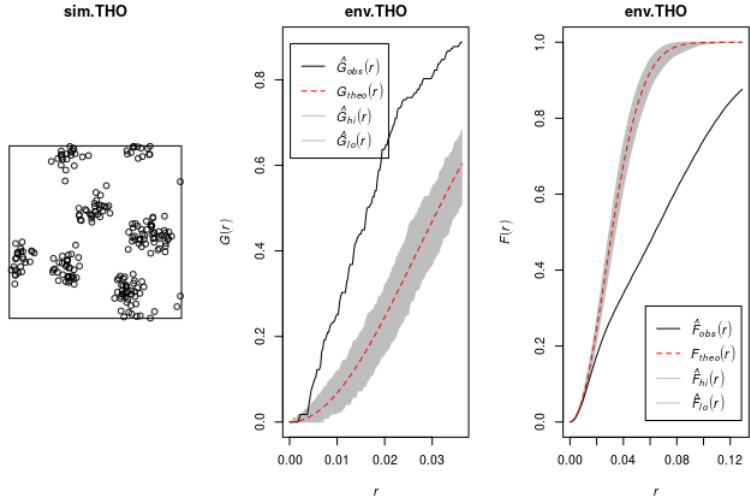
Para mostrar la utilidad de estas métricas, tomaremos algunos de los procesos vistos previamente, y calcularemos la función G en cada uno de ellos.

Para el proceso de Poisson homogéneo obtenemos:



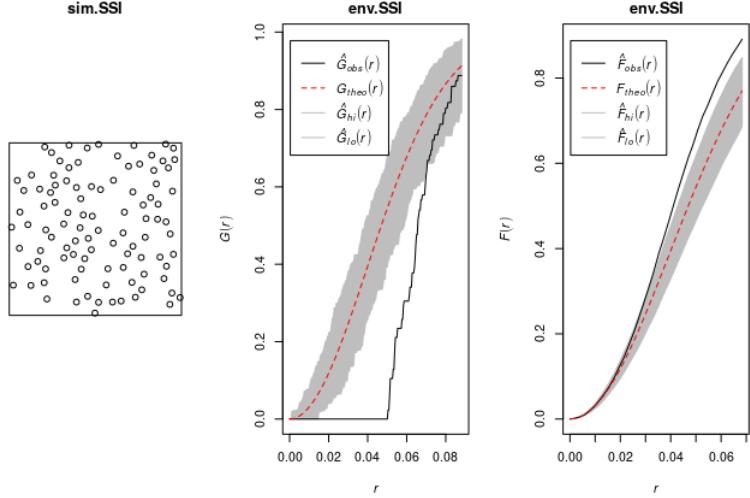
Puede verse que las funciones empíricas F y G (curvas sólidas) se mantienen bien adentro de la envolvente esperada para un proceso CSR.

Para el proceso de Thomas obtenemos:



En este caso, como es de esperar, la función G se posiciona por encima de la envolvente, pues las distancias chicas a los vecinos más cercanos son fáciles de alcanzar desde los mismos eventos, ya que se hallan fuertemente concentrados. Inversamente, la función F se posiciona por debajo de la envolvente, pues las distancias chicas a los vecinos más cercanos son menos probables de alcanzar desde puntos arbitrarios dentro de la región de análisis. Sólo cuando los puntos de la región caen cerca o dentro de los clusters, se obtienen proporciones importantes de distancias chicas.

Para el proceso de Inhibición Simple obtenemos:



Este caso es inverso al anterior, pues la función G acusa una nula proporción de eventos muy cercanos entre sí, por la inhibición. Mientras que la función F muestra proporciones altas de distancias chicas a eventos, pues la regularidad del patrón puntual le asegura a cada punto arbitrario dentro de la región un evento vecino cercano.

Nótese también que en los dos últimos casos la función F es mucho más suave que la función G , manteniendo una relación inversa con respecto al benchmark de CSR.

La Función K de Ripley

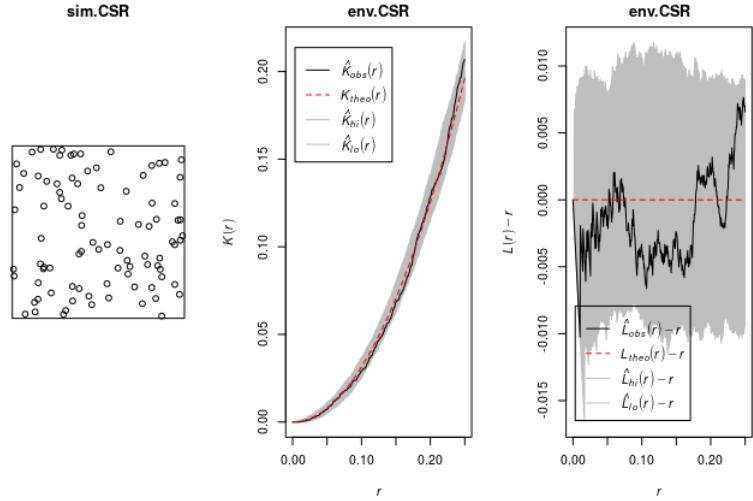
La métrica más difundida para cuantificar el alejamiento de un patrón de eventos en relación a un proceso CSR es la función K de Ripley. Esta función $K(r)$ mide, salvo una constante, la proporción de distancias entre eventos d_{ij} que son menores a r . O sea

$$K(r) = \frac{\|A\|}{n(n-1)} \sum_{i \neq j} I(d_{ij} < r) = \lambda^{-1} \sum_{i \neq j} \frac{I(d_{ij} < r)}{n-1}$$

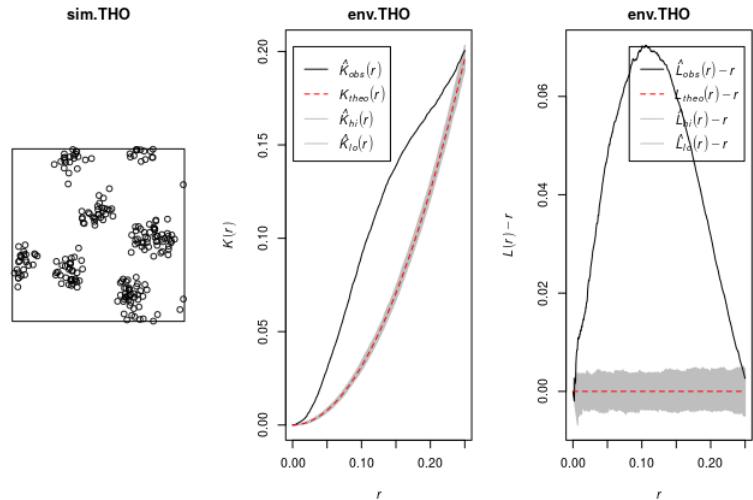
Para un proceso CSR, la función equivale a $K(r) = \pi r^2$. Para simplificar la comparación, se aplica una simple transformación a esta función, obteniéndose la función $L(r) = \sqrt{\frac{K(r)}{\pi}}$. Con esta última función se suele graficar $r - L(r)$ versus r , y buscar alejamientos del valor 0.

Nuevamente, observemos para los procesos definidos previamente el comportamiento de la función.

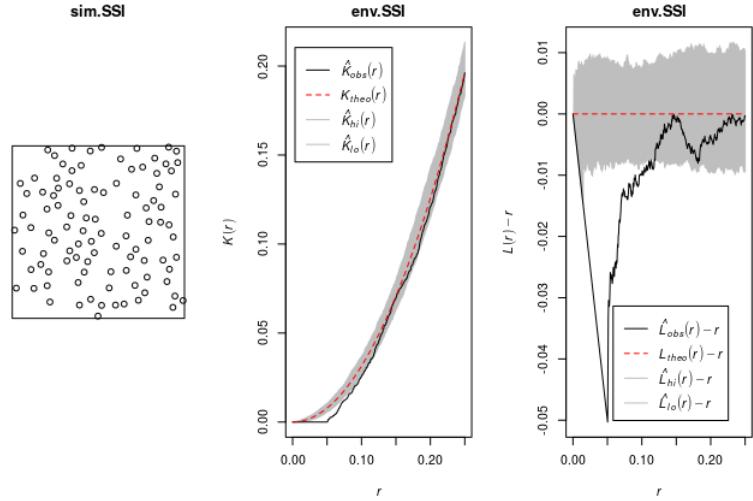
Para el proceso de Poisson homogéneo obtenemos:



Para el proceso de Thomas obtenemos:



Para el proceso de Inhibición Simple obtenemos:



Nótese que en el caso de procesos de concentración, la función empírica queda por encima de la envolvente, ya que preponderan las distancias pequeñas. En tanto que en las situaciones de regularidad, las distancias muy cortas están subrepresentadas, como en el proceso de inhibición, en el que se excluyen vecinos muy cercanos.

Estimación de la Intensidad del Proceso

Si tuvieramos la sospecha que el proceso generador de los eventos es No homogéneo, como es de esperar en la mayoría de los casos, surge la necesidad de inferir la función de intensidad subyacente al proceso.

El enfoque más simple proviene de suponer que el proceso generador es un proceso de Poisson inhomogéneo. Bajo este supuesto es fácil ver que (Diggle 8.2) la log-verosimilitud del proceso puede escribirse como

$$L(\lambda) = \sum_{i=1}^n \log(\lambda(x_i)) - \int_A \lambda(x) dx$$

donde x_i es la localización del evento i -ésimo, A es la región de análisis, y $\lambda(\cdot)$ es la función de intensidad desconocida que debemos modelar. Usualmente se propondrá alguna función paramétrica para $\lambda_\beta(x)$, como por ejemplo algún polinomio de bajo grado con parámetros β . En este caso el enfoque es completamente paramétrico, y la estimación consiste en buscar aquel valor de los parámetros que minimicen la expresión $L_\beta(x_1, \dots, x_n)$. Un aspecto interesante de esta expresión, es que resalta claramente el trade-off sesgo-varianza inherente al proceso de estimación. El primer término de la expresión se encarga de “ajustar” la función de intensidad a la muestra (reducción de sesgo), en tanto que el segundo término penaliza globalmente el exceso de ajuste (reducción de varianza).

Desde los enfoques no paramétricos, la técnica más difundida de estimación de la intensidad es la de KDE (Kernel Density Estimation). El estimador se define como

$$\hat{\lambda}(x) = \frac{1}{h^2} \sum_{i=1}^n K\left(\frac{\|x - x_i\|}{h}\right)$$

donde x es el punto de estimación, h es un escalar que sirve para regular la granularidad de la estimación (ventana o bandwidth), y K es una función positiva de nucleo que sirve para ponderar las observaciones, dándole mayor importancia a las más cercanas al punto donde se está estimando la intensidad.

Este estimador de la intensidad, es equivalente el estimador de densidad por nucleos que se usa para estimar densidades. La única diferencia es la falta de la constante $\frac{1}{n}$ que asegura que sea una densidad (que integre 1).

El kernel más utilizado es el gausiano. Otro kernel muy usual para el cálculo de intensidades en estadística espacial es el kernel cuadrático:

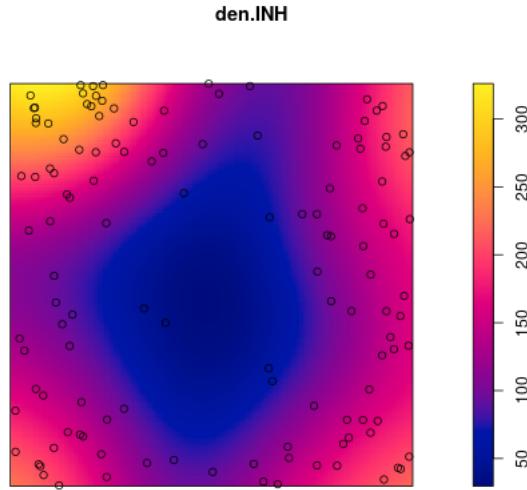
$$K(u) = \begin{cases} \frac{3}{\pi} \left(1 - \|u\|^2\right)^2 & u \in (-1, 1) \\ 0 & \text{cc} \end{cases}$$

Veamos ahora como se comporta el estimador de intensidad por nucleos en algunas de las realizaciones de procesos vistas antes. Empecemos con el proceso de Poisson Inhomogeneo.

Estimamos la intensidad del patrón de puntos que fue generado con función de intensidad:

$$\lambda(x,y) = 800 \left(x - \frac{1}{2} \right)^2 + 800 \left(y - \frac{1}{2} \right)^2$$

en la que la intensidad se minimiza en el centro de la ventana $[0, 1] \times [0, 1]$, con un valor de 0, y se maximiza hacia los bordes, en donde alcanza valores que oscilan entre 200 y 400.(en las esquinas). El siguiente gráfico muestra el resultado de aplicar el estimador KDE con un núcleo normal y un valor de ventana de 0.15.



Se aprecia que la estimación es bastante precisa. Pero de donde sacamos este valor de ventana $h = 0.15$? Una modalidad usual y efectiva es calcular el h óptimo por validación cruzada de un elemento (LOOCV). Se aproxima el error (esperado) entre la intensidad estimada y la real para diferentes valores de h y se elige el que arroje un menor error.

La idea es hallar un valor de h que minimice el siguiente error

$$Error(h) = \int \left(\widehat{\lambda}_h(x) - \lambda(x) \right)^2 dx$$

que es equivalente a

$$Error(h) = \int \widehat{\lambda}_h^2(x) dx - 2 \int \widehat{\lambda}_h(x) \lambda(x) dx + \int \lambda^2(x) dx$$

como el último término no depende de h podemos eliminarlo, y nos queda

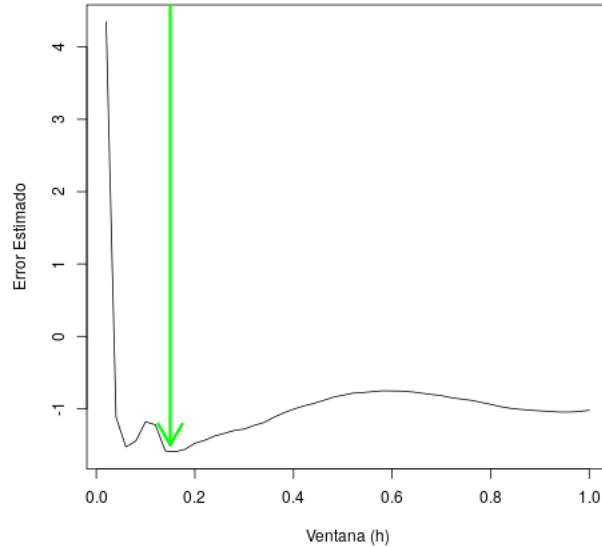
$$Error^*(h) = \int \widehat{\lambda}_h^2(x) dx - 2 \int \widehat{\lambda}_h(x) \lambda(x) dx$$

que puede ser aproximado por

$$\widehat{Error}^*(h) = \int \widehat{\lambda}_h^2(x) dx - 2 \sum_i \widehat{\lambda}_h^{-i}(x_i) \frac{1}{n}$$

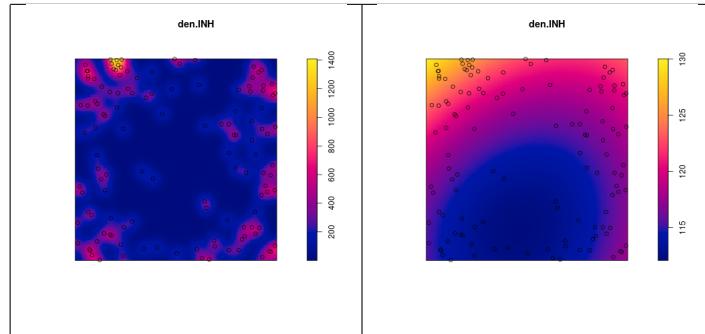
donde $\widehat{\lambda}_h^{-i}(x_i)$ es la estimación (sin la observación i -ésima) de la intensidad evaluada en la observación i .

En nuestro caso, el gráfico de la estimación del error $\widehat{Error}^*(h)$ por LOOCV es (la flecha verde señala el valor $h = 0.15$)

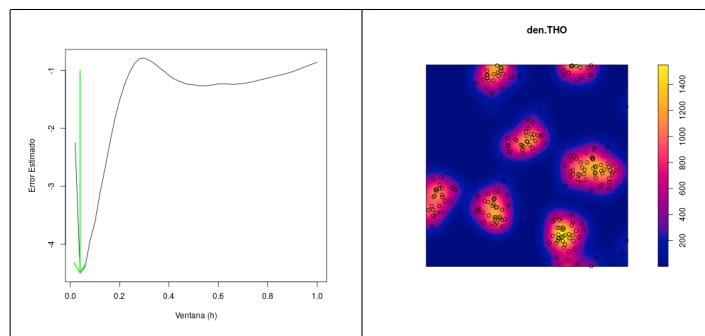


Claramente, ventanas más grandes (y demasiado chicas) aumentarían el error.

Veamos dos situaciones desfavorables, una con una ventana muy chica (panel izquierdo, $h = 0.03$), y la otra con un valor de ventana exageradamente grande (panel derecho, $h = 0.6$).



Veamos ahora que sucede con la estimación de la intensidad del patrón de puntos generado por el proceso de Thomas.



Efectivamente, el estimador encuentra zonas de alta densidad de concentración de eventos. Sin embargo, como ha sido mencionado numerosas veces, con una sola realización del proceso no podemos saber si es la función de intensidad la que genera la concentración, o un efecto de interacción entre eventos, como en este caso.

Geoestadística: Interpolación y Predicción Espacial

La interpolación y la predicción espacial suelen considerarse como técnicas esenciales del campo de estudio denominado **geoestadística**. En geoestadística, el objetivo fundamental es el de analizar el comportamiento de una o varias variables Z que son medidas en diferentes puntos del espacio s , es decir $Z(s)$. Sin embargo, en este caso, el interés no está centrado en el análisis de las posiciones (s) de las mediciones, sino en las mediciones (Z). Así, las propiedades de heterogeneidad, concentración, o regularidad de las ubicaciones no son un objeto de estudio en sí mismo, en tanto que sí son un atributo insoslayable del estudio de Z . El problema básico de la geoestadística es el de pronosticar el valor no observado de $Z(s_0)$ habiendo observado $Z(s_1), Z(s_2) \dots Z(s_N)$, cuando las ubicaciones espaciales s_1, s_2, \dots, s_N se hallan cerca de s_0 , como se muestra en la próxima figura



En este ejemplo, donde se cuenta con precios de venta de propiedades en la ciudad de Buenos Aires, se busca predecir el precio de una nueva propiedad en función del conocimiento de los precios y ubicaciones de otras propiedades cercanas.

Existen varias herramientas para realizar esta tarea, siendo tres de las más difundidas:

- Inverse Distance Weighting (IDW)
- Geographically Weighted Regression (GWR)
- Kriging

En lo que sigue resumiremos brevemente estas herramientas.

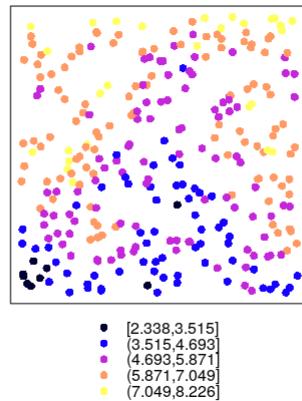
Pero antes vamos a generar datos espaciales simulados que nos servirán para aplicar las técnicas, y adicionalmente para comprender las características esenciales del fenómeno que queremos modelar.

Simulación Espacial: Proceso Marcado Autocorrelacionado

Vamos a simular $N = 300$ datos de un proceso espacial $Y(x_1, x_2)$ continuo en el soporte $x_1 = [0, 1] \times x_2 = [0, 1]$ con las siguientes características:

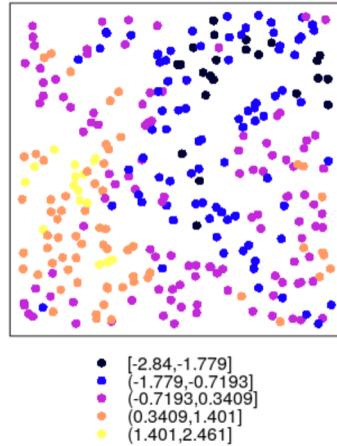
- Proceso marcado, con una variable de marca continua Y cuyas realizaciones tomarán valores entre 2.3 y 8.3.
- Una tendencia espacial lineal determinística dada por $E(Y) = 3 + 2 * x_1 + 4 * x_2$
- Una estructura de autocorrelación espacial positiva de los errores ε , por lo que observaciones cercanas tenderán a mostrar valores parecidos, más allá de su componente determinística compartida. Específicamente, la estructura de correlaciones de los errores obedece a $Cor(\varepsilon_i, \varepsilon_j) \propto Exp(-5 \| \mathbf{X}_i - \mathbf{X}_j \|)$.

Veamos la realización del proceso con la que trabajaremos de ahora en más.



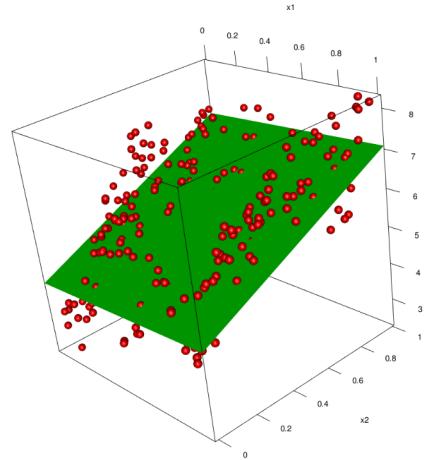
Observese que los valores de marca aumentan en gran medida de sur a norte (x_2), y en menor medida de oeste a este (x_1).

También es interesante mostrar el mapa de los errores, es decir, el mapa de los valores observados (realizados) deducida la tendencia espacial real (poblacional).



Nótese que los errores se hallan claramente autocorrelacionados. De hecho, los colores no se hallan distribuidos aleatoriamente en el mapa, mas bien se encuentran aglomerados espacialmente en relación a la similaridad de sus valores.

Todo esto puede verse simultaneamente en el siguiente gráfico de perspectiva



En el gráfico puede apreciarse el plano inclinado (la tendencia espacial determinística), y la autocorrelación, evidente por el agrupamiento de las observaciones por debajo (y encima) del plano (fíjese el faltante de observaciones sobre el plano en la parte central del mismo).

A los fines puramente descriptivos, una herramienta muy utilizada para cuantificar la componente de autocorrelación espacial es el Indice I de Moran.

El Índice de Moran

Este índice mide la autocorrelación espacial de una variable X con mediciones en el espacio. Se define como

$$I = \frac{n}{\sum_{i,j} W_{ij}} \frac{\sum_{i,j} W_{ij} (X_i - \bar{X}) (X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$$

donde:

X_i es la medición i-ésima de la variable de análisis

\bar{X} es la media global de la variable

$W_{i,j}$ es un ponderador relacionado inversamente con la distancia

Bajo la hipótesis nula de no autocorrelación, este índice tiene un valor esperado de $-1/(n-1)$, muy cercano a cero cuando n es grande.

Vale la pena reescribir al índice descomponiéndolo en n términos aditivos, de la siguiente manera

$$\begin{aligned} I &= \frac{n}{\left(\sum_{i,j} W_{ij}\right) \sum_i (X_i - \bar{X})^2} \sum_i (X_i - \bar{X}) \sum_j W_{ij} (X_j - \bar{X}) = \\ &= \sum_i \left(\frac{n (X_i - \bar{X})}{\left(\sum_{i,j} W_{ij}\right) \sum_i (X_i - \bar{X})^2} \sum_j W_{ij} (X_j - \bar{X}) \right) = \\ &= \sum_i I_i \end{aligned}$$

A cada uno de estos n términos (I_i) se los llama “Local Moran”, y miden, centrados en cada observación, el grado de correlación que esa observación tiene con su entorno. El índice local de Moran queda así

$$I_i = \frac{n (X_i - \bar{X})}{\left(\sum_{i,j} W_{ij}\right) \sum_i (X_i - \bar{X})^2} \sum_j W_{ij} (X_j - \bar{X})$$

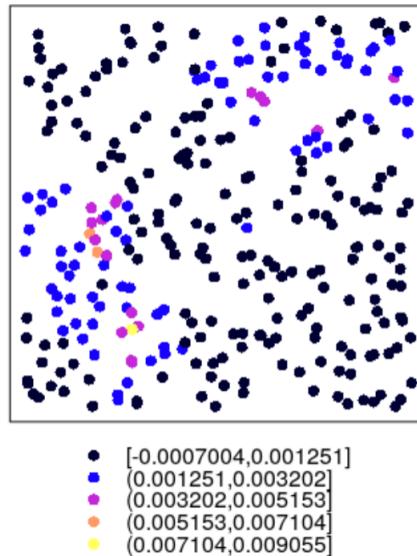
Es claro que, salvo constantes, el índice capta la relación entre el desvío de la observación i-ésima con respecto a la media ($X_i - \bar{X}$), y los desvíos de las observaciones cercanas $\sum_j W_{ij} (X_j - \bar{X})$, si pensamos en ponderadores que dan valor 1 a las observaciones cercanas y 0 a las lejanas. En situaciones de autocorrelación positiva, el índice debiera dar valores positivos.

Es importante considerar que si la variable X posee una tendencia espacial (proceso no homogéneo), el índice de Moran arrojará, trivialmente valores significativos y positivos, aún cuando no exista autocorrelación espacial. ¿Por qué

? Porque las observaciones de X que se encuentran en zonas de alta tendencia tendrán valores por sobre su media global, pero las observaciones cercanas también, ya que comparten la misma tendencia ! Estaríamos confundiendo inhomogeneidad con autocorrelación.

En el caso de nuestros datos simulados sabemos que existe una tendencia, por lo que aplicaremos el Índice de Moran a los residuos de los datos. El índice arroja un valor de 0.25, muy alejado del valor esperado de 0.003 bajo la hipótesis nula de no autocorrelación.

Si graficamos los valores de Moran locales para nuestros datos, podemos ver el siguiente mapa

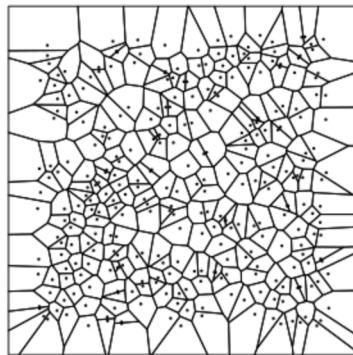


Por supuesto que los valores están agrupados. Lo revelador de este tipo de mapas es la evaluación de la localización de los valores más elevados. En este caso, la diagonal suroeste-noreste parece ser la zona de mayor autocorrelación espacial positiva. Mas adelante veremos este análisis aplicado a datos reales.

Celdas de Voronoi o Nearest Neighbor Interpolation

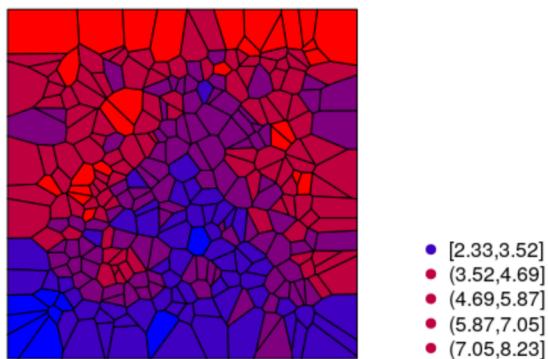
El método de interpolación más sencillo que puede pensarse es aquel que predice el valor de cualquier punto del espacio $X_0 = X(s_0)$ con el valor observado más cercano a ese nuevo punto, es decir $\widehat{X}_0 = \{X_i : s_i = \text{Mink} \| s_k - s_0 \| \}$. Esto es equivalente a particionar la región de estudio en sus celdas de Voronoi, y asignar a cada celda el valor de la observación contenida en la celda.

La partición de Voronoi (Voronoi tessellation) de nuestros datos es



Cada celda posee una única observación en su interior, con fronteras formadas por las bisectrices de equidistancia entre pares de observaciones.

La interpolación de Voronoi aplicada a nuestros datos simulados es



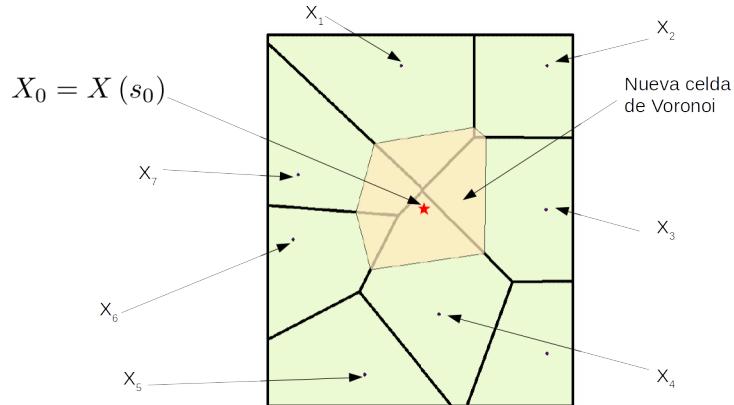
Claramente, el resultado es un campo no continuo con alta chance de sobreajustar (a la muestra) el fenómeno (proceso) subyacente.

Una inmensa mejora a esta técnica, llamada Natural Neighbor Interpolation, predice el valor de cualquier punto del espacio $X_0 = X(s_0)$ con el promedio ponderado de los valores de observaciones cuyas celdas de Voronoi intersectan la “nueva” celda de Voronoi del punto a ser predicho, con ponderadores proporcionales a las áreas de intersección de las celdas con esta nueva celda. Es decir

$$\widehat{X}_0 = \sum_i \frac{A(X_i)}{A(X_0)} X_i$$

donde $A(X_i)$ es el área de intersección de la celda de Voronoi de X_i con la nueva celda de X_0 , y $A(X_0)$ es el área de la celda del punto a ser predicho.

El próximo gráfico muestra un ejemplo de este método

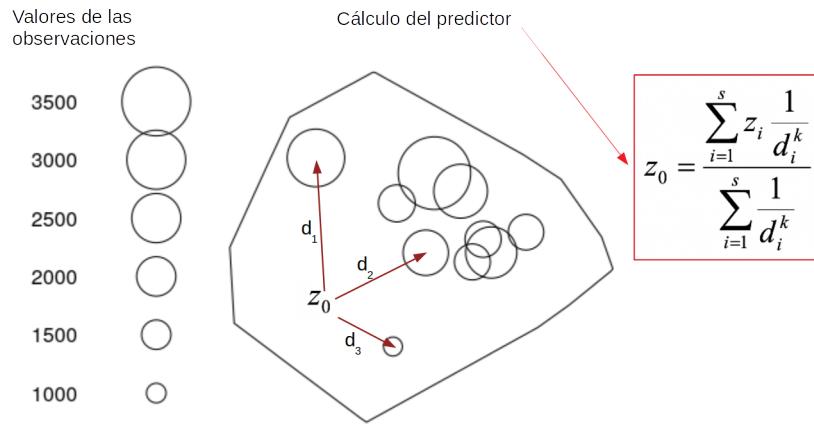


En color naranja se representa el área de la celda de Voronoi del punto a ser predicho X_0 , que genera intersección con 6 observaciones. La observación que más (menos) influye en la predicción es X_4 (X_2), mientras que la observación X_5 no contribuye en nada a la estimación.

A diferencia de la interpolación por el vecino más cercano, esta técnica produce una superficie suave de predicción, conservando la estimación en los puntos de muestra coincidentes con los valores observados.

Inverse Distance Weighting (IDW)

Esta es otra de las técnicas más sencillas de interpolación, y consiste simplemente en predecir un nuevo valor mediante el promedio ponderado de los valores cercanos, haciendo que las ponderaciones sean inversamente proporcionales a sus respectivas distancias en relación a la posición del valor a ser predicho. La intuición del método y la regla de cálculo pueden verse en la siguiente figura.

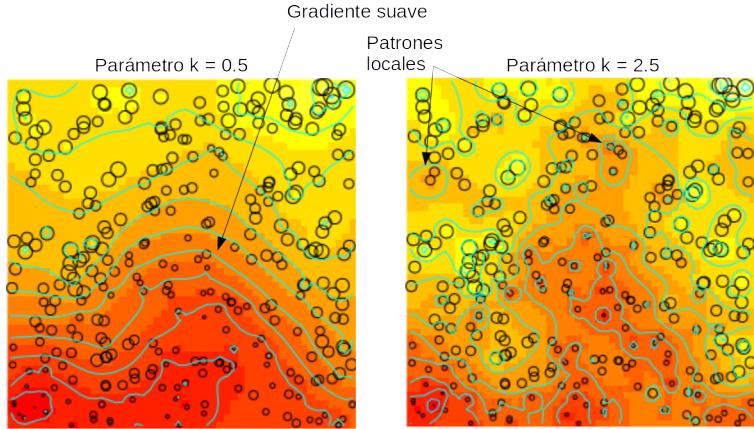


De lo que se desprende de la figura, si tuvieramos que predecir a $Z(s_0)$ en función de las 3 observaciones a las que apuntan las flechas, parece razonable que la predicción se vea más influenciada por la medición de menor valor (círculo pequeño al sur-este), y menos por la medición de mayor valor (círculo grande al norte).

Esta técnica provee, como resultado de su aplicación, un campo de interpolación continua (suave) para la región de análisis, excepto para las ubicaciones de las mismas observaciones, en las que el predictor no está definido (división por cero !). Para los casos particulares de evaluación del predictor en puntos observados, para los cuales la distancia es cero, se define el valor predicho igual al observado en esos puntos, para solucionar el problema de indefinición que generan las distancias iguales a cero.

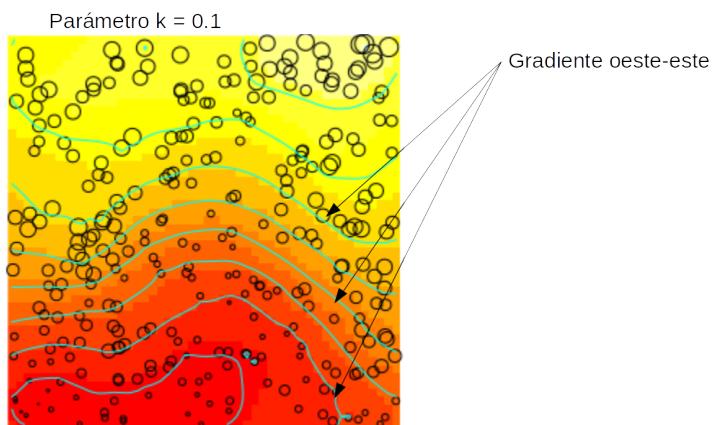
El parámetro que regula el nivel de complejidad de la interpolación es el grado del exponente (k) que afecta a la distancia, el cual deberá tomar valores estrictamente positivos. Un incremento del valor de este parámetro produce una estimación más “local”, ya que las observaciones más cercanas recibirán proporcionalmente mucho más peso. Contrariamente, reducciones del valor del parámetro producirán interpolaciones más “globales”, captando con mayor facilidad tendencias espaciales, en detrimento de patrones puntuales.

Mostramos ahora la aplicación ejemplo con $N = 300$ observaciones para las cuales los valores de medición aumentan (círculos grandes) de sur a norte, y en menor medida de oeste a este.



Se aprecia como el valor de $k = 2.5$ detecta patrones locales, mientras que el valor del parámetro $k = 0.5$ sólo capta la tendencia más importante de orientación sur-norte.

Es interesante notar que la tendencia oeste-este no se puede observar, ni siquiera con un valor reducido del parámetro de $k = 0.5$. Para poder rescatar esa tendencia debieramos achicar más aún el valor del parámetro, llevandolo a $k = 0.1$. como se muestra en este último gráfico.



La pregunta de siempre es: ¿Como se determina el valor óptimo del parámetro? La respuesta de siempre es: Validación Cruzada ??? No necesariamente. Puede no existir un valor óptimo de parámetro. Puede ser revelador generar varios análisis con diferentes valores, y comparar los resultados. La generación

de una secuencia de análisis de este tipo podría enseñarnos posibles tendencias globales, e interesantes patrones locales, como en el ejemplo recién presentado.

Kriging

El Kriging es una técnica de predicción espacial para procesos marcados que se originó en la industria de la minería. La idea principal es la de descomponer a un proceso espacial en dos componentes. Un primer componente de variación espacial de gran alcance o de alcance global $\mu(S)$. Un segundo componente de variación espacial de corto alcance o de alcance local $\epsilon_\Sigma(S)$. Así el fenómeno se puede expresar

$$Z(S) = \mu(S) + \epsilon_\Sigma(S)$$

donde

la componente $\mu(S)$ suele incorporar efectos de predictoras, y la componente $\epsilon_\Sigma(S)$ se encarga de modelar los efectos de la autocorrelación espacial, modelados a través de la matriz de covarianzas de los errores (Σ).

A grandes rasgos existen tres tipos de kriging distintos, cada uno obedeciendo a diferentes supuestos sobre la primer componente.

Lo que se conoce con **Simple Kriging** asume una intensidad media μ constante y conocida, o sea

$$Z(S) = \mu + \epsilon_\Sigma(S)$$

En el caso que podamos suponer una intensidad media constante pero desconocida m , tendremos **Ordinary Kriging**

$$Z(S) = m + \epsilon_\Sigma(S)$$

Finalmente, el modelo más útil y realista, llamado **Universal Kriging**, consiste en suponer una intensidad media no constante, dependiente (potencialmente) de predictoras X que pueden tener una componente espacial.

$$Z(S) = X(S)\beta + \epsilon_\Sigma(S)$$

En este último caso, la predicción final requerirá la estimación de los parámetros desconocidos β , así como de la estructura de correlación espacial Σ .

El Modelo de Kriging como Predictor

El modelo de kriging puede ser considerado un caso particular del modelo que motiva a la técnica Generalized Least Squares (GLS), que expande al modelo lineal clásico al caso en el que los errores NO son independientes. En nuestro caso, es de esperar que los errores muestren una estructura de autocorrelación (a través de Σ) dependiente de la distancia de las localizaciones de las observaciones.

La predicción puntual que plantea kriging para $Z_0 = Z(s_0)$ sucede escribirse así

$$\hat{Z}_0 = X_0\hat{\beta} + vV^{-1}\left(\mathbf{z} - X\hat{\beta}\right)$$

donde

\hat{Z}_0 es la respuesta a predecir para la la posición s_0

Z es el vector de valores observados de las muestras localizadas en $s_1 \dots s_n$

X_0 es el vector de predictoras asociadas a la respuesta $Z_0 = Z(s_0)$

$\hat{\beta}$ es el vector estimado de coeficientes que capta el efecto de las predictoras

v es el vector de covarianzas entre los errores de Z_0 y Z , que depende esencialmente de las distancias de s_0 al resto de las observaciones $s_1 \dots s_n$

V es la estimación de la matriz de covarianzas de los errores Σ

X es la matriz de datos

De la expresión surge claramente que la predicción contempla dos efectos, el efecto de las predictoras en el punto de interés $X_0\hat{\beta}$ y el efecto que los residuos $(Z - X\hat{\beta})$ tienen, dependiendo de las distancias al punto en cuestión s_0 , esto último captado por los pesos de kriging vV^{-1} . Nótese que ante la falta de autocorrelación espacial ($v = \mathbf{0}$) el segundo término se anula, y el predictor coincide con un simple predictor lineal.

Asimismo, en el caso general en el que existe autocorrelación espacial. el estimador del vector de coeficientes es

$$\hat{\beta} = (XV^{-1}X^T)X^TV^{-1}Y$$

que es el estimador GLS considerando la (estimacion) de la estructura de correlación de los errores dependiente del espacio.

La estimación V de la matriz de covarianzas de los errores Σ , no es una tarea trivial, pues la misma contempla (potencialmente) todas las covarianzas de los pares de observaciones (en realidad de sus errores) de la muestra. Es casi imposible estimar esta matriz sin imponer una fuerte estructura a la misma. El planteo que propone kriging es el de postular una relación directa entre esta matriz y la dependencia espacial de los residuos en términos de sus distancias espaciales. A esta dependencia espacial de la variabilidad en función de la distancia se la denomina **Variograma**.

El Variograma

Sea un proceso espacial $Z(\mathbf{s})$. Tomemos dos elementos del proceso, ubicados en \mathbf{s}_1 y en \mathbf{s}_2 . El variograma mide la dependencia espacial entre estos elementos, de la siguiente manera:

$$\begin{aligned}\gamma(\mathbf{s}_1, \mathbf{s}_2) &= \text{Var}(Z(\mathbf{s}_1) - Z(\mathbf{s}_2)) = \\ &= E(Z(\mathbf{s}_1) + E(Z(\mathbf{s}_2)) - Z(\mathbf{s}_2) - E(Z(\mathbf{s}_1)))^2\end{aligned}$$

Si el proceso es homogéneo entonces $E(Z(\mathbf{s}_1)) = E(Z(\mathbf{s}_2))$ por lo que

$$\gamma(\mathbf{s}_1, \mathbf{s}_2) = E(Z(\mathbf{s}_1) - Z(\mathbf{s}_2))^2$$

Si el proceso es estacionario e isotrópico entonces podemos eliminar los índices y pensar que $Z(\mathbf{s}_1) = Z(\mathbf{s})$ y $Z(\mathbf{s}_2) = Z(\mathbf{s} + \mathbf{h})$, por lo que

$$\gamma(\mathbf{s}_1, \mathbf{s}_2) = E(Z(\mathbf{s}) - Z(\mathbf{s} + \mathbf{h}))^2 = \gamma(\mathbf{h})$$

Que es una función que sólo depende del vector de separación \mathbf{h} . Es claro que $\gamma(\mathbf{h}) \geq 0 \forall \mathbf{h}$.

Pero hay que tener mucho cuidado ! Pues si la variable Z muestra una tendencia espacial importante, o está influenciada por predictoras con un fuerte componente espacial, el variograma se verá afectado por estas relaciones, confundiendo efectos locales que dependen de la distancia, con efectos globales que dependen del espacio a nivel global.

Así, en estos casos más generales, que modelamos con

$$Z(S) = \mu(S) + \varepsilon_{\Sigma}(S)$$

suele definirse el variograma en función de los errores ε , por lo que expresamos al variograma de la siguiente manera

$$\gamma(\mathbf{h}) = E(\varepsilon(\mathbf{s}) - \varepsilon(\mathbf{s} + \mathbf{h}))^2$$

donde $\varepsilon(\mathbf{s})$ y $\varepsilon(\mathbf{s} + \mathbf{h})$ son dos errores distintos cualesquiera, asociados a las observaciones, separados por el vector \mathbf{h} . Estrictamente, esta definición obedece al **semivariograma**, que es la mitad del variograma, sin embargo de ahora en más omitiremos esta distinción y lo llamaremos variograma, después de todo sólo difieren en una constante (1/2).

Varios aspectos fundamentales del variograma deben ser considerados:

- El varograma, así definido, sólo depende del vector de distancia \mathbf{h} entre las observaciones, NO del lugar del espacio en cuestión (\mathbf{s}), por lo que se supone que el proceso de medición, una vez descontado los efectos de las predictoras, es **homogéneo**.
- Para $\mathbf{h} = \mathbf{0}$ la función debiera tomar el valor 0, pues $E(\varepsilon(\mathbf{s}) - \varepsilon(\mathbf{s}))^2 = 0$. Sin embargo, en la práctica sólo pueden considerarse mediciones del variograma superiores a una distancia mínima, para la cual se obtienen valores de variograma también mínimos. A este valor inicial y mínimo del variograma se lo llama “**nugget**”.
- Como los errores tienen media 0, por definición $\gamma(\mathbf{h}) = Var(\varepsilon(\mathbf{s}) - \varepsilon(\mathbf{s} + \mathbf{h}))$, es decir que el variograma no es otra cosa que la varianza de la discrepancia entre los errores separados por \mathbf{h} .
- Puede verse que $\gamma(\mathbf{h}) = 2Var(\varepsilon(\mathbf{s})) - 2Cov(\varepsilon(\mathbf{s}), \varepsilon(\mathbf{s} + \mathbf{h}))$, lo que muestra que el variograma, para un valor de \mathbf{h} dado, depende positivamente de la variabilidad del proceso, y negativamente de la covarianza del mismo.
- A partir de una cierta distancia $\|\mathbf{h}\| > r$ (a este r se lo llama “**range**”), debiera suceder que $Cov(\varepsilon(\mathbf{s}), \varepsilon(\mathbf{s} + \mathbf{h})) = 0$, por lo que $\gamma(\mathbf{h}) = 2Var(\varepsilon(\mathbf{s}))$,

y el variograma debiera volverse una constante que sólo dependa de la varianza de ε , es decir de $\sigma^2(\varepsilon)$. A este valor constante se lo llama “**sill**”.

Usualmente el variograma se modela paramétricamente mediante algunos modelos relativamente sencillos que dependen de pocos parámetros. La elección del tipo de modelo se realiza en general a partir de versiones empíricas del variograma. La versión empírica del variograma más usual consiste en armar buckets de distancias, y para cada bucket calcular el promedio de los cuadrados de las diferencias entre residuos, para aquellas observaciones cuyas distancias caigan en ese bucket.

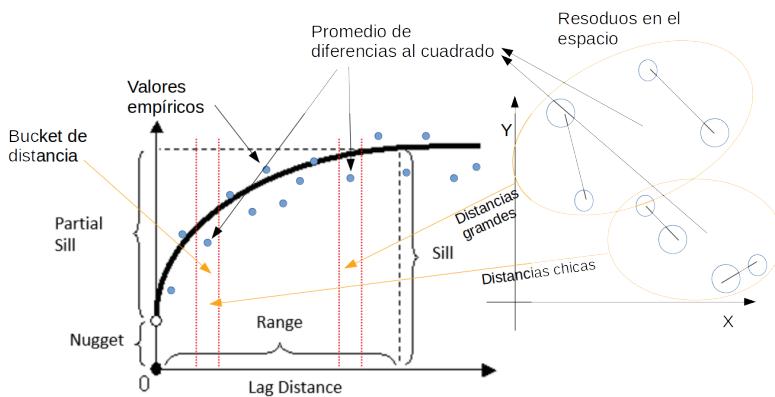
Uno de los modelos paramétricos más comunes de variograma es el exponencial, que se define así

$$\gamma(h) = (\text{sill} - \text{nug}) \left(1 - \exp\left(-\frac{h}{a * \text{range}}\right) \right) + \text{nug} * 1(h > 0)$$

Este modelo depende de los parámetros:

- nug: nugget, el mínimo valor del variograma cerca de la distancia cero
- range: rango, el lapso de distancia hasta que el variograma alcanza un valor constante
- sill: el valor constante asintótico del variograma
- a: constante de escala dependiente de la distancia, usualmente igual a 1.

Veamos un ejemplo de variograma empírico (puntos azules), con la superposición de un modelo exponencial (curva negra).



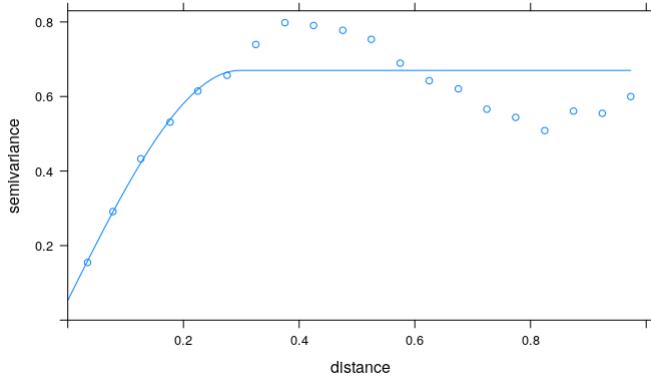
La figura muestra (al menos lo intenta) el proceso de armado del variograma empírico, para el cual se determinan los buckets de distancias (líneas punteadas naranjas), y se calculan los promedios de las diferencias al cuadrado de los

valores entre los pares de residuos que quedan a distancias comprendidas en esos buckets (elipses naranjas de la derecha). Finalmente, los parámetros del modelo exponencial son ajustados a los valores del variograma empírico, para obtener así la estimación de la matriz de covarianzas V que dependerá sólo de las distancias entre observaciones a través del modelo exponencial.

Ejemplo de Kriging a Datos Simulados

Trabajaremos con los mismos $N = 300$ datos simulados que utilizamos para el método de interpolación IWD.

Empecemos mostrando para estos datos el variograma empírico junto con el ajuste de un modelo exponencial.



El ajuste no es muy bueno porque, en este caso, se aprecia un efecto llamado “hole-effect”, producto de un comportamiento cíclico que suele hallarse en fenómenos espaciales autocorrelacionados cuando los residuos se agrupan en valores similares a lo largo del espacio a distancias medianas y grandes. Esto produce que a lags de distancias grandes disminuya el valor del variograma, pues los residuos muestran comportamientos similares debido al agrupamiento de los mismos. De todas maneras no ajustaremos este efecto (se puede, usando modelos apropiados que suman componentes sinusoidales al variograma teórico), pues lo más importante del modelado consiste en considerar el primer tramo creciente, en el que se capta la fuerte autocorrelación espacial del fenómeno.

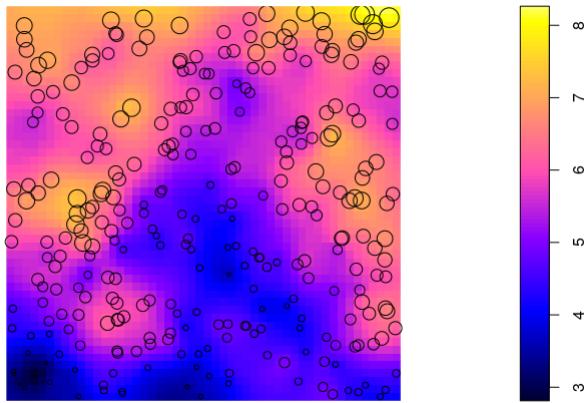
Ajustaremos ahora un modelo que involucre una componente determinística y una estocástica

$$Z(S) = X(S)\beta + \varepsilon_{\Sigma}(S)$$

En nuestro caso el modelo propuesto presentará una tendencia espacial lineal (como la verdadera, con la que se generaron los datos), y la componente de autocorrelación modelada con el modelo de variograma exponencial

$$Z(S) = \mu + x_1\beta_1 + x_2\beta_2 + \varepsilon_{V_{Exp}}(S)$$

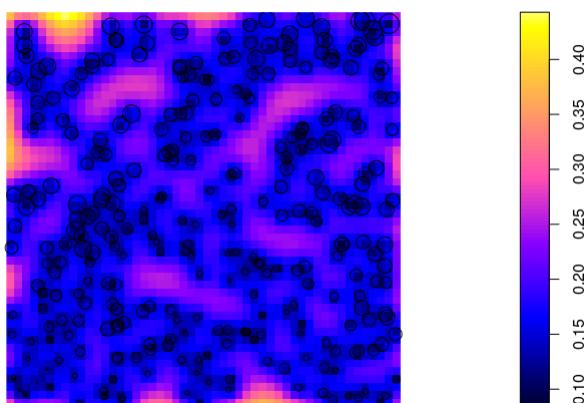
Veamos ahora el campo predicho por este modelo



Es interesante notar que el resultado es muy parecido al de la interpolación mediante IDW (compare ambos gráficos). Pero entonces, cual es la ventaja de kriging ? Varias ventajas:

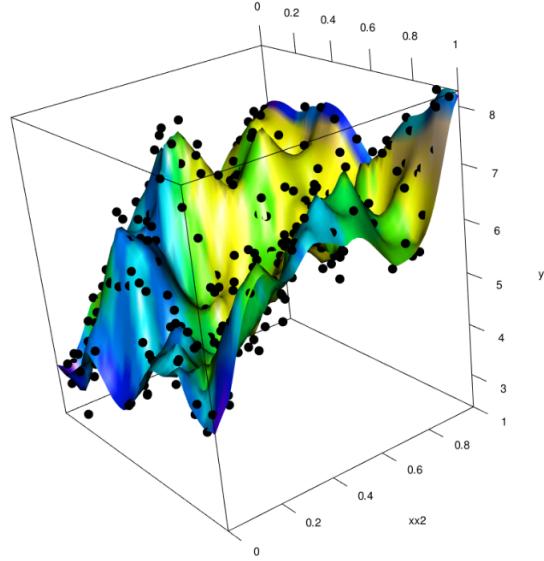
- La posibilidad de incorporar (controlar por) covariables predictoras, sean espaciales o no.
- La estimación del efecto de las predictoras que conforman la componente determinística del modelo.
- La obtención de una medida creíble de incertidumbre de la predicción, proveniente de los desvíos estandar calculados por el modelo.

Veamos, en relación a el último punto, el mapa de incertidumbre (varianza) en las predicciones.



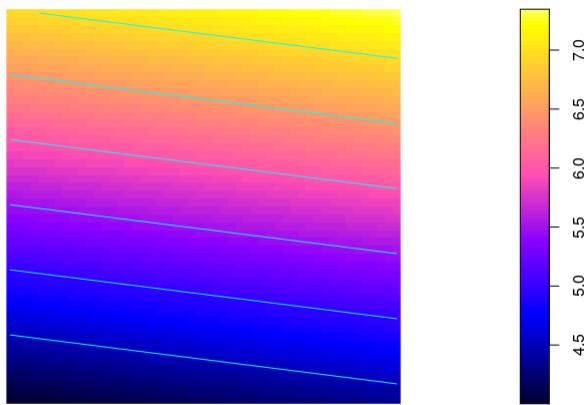
Como es razonable, las mayores varianzas se encuentran en los bordes de la región de análisis y en las zonas de menor concentración de eventos.

Otra forma de ver el ajuste es un gráfico de perspectiva en que se muestra el ajuste de kriging junto con las observaciones.



Veáse que la interpolación es una mezcla del efecto de tendencia espacial (creciente de izquierda a derecha y de abajo hacia arriba), junto con el efecto local producto del modelado de la autocorrelación (picos y valles por sobre la tendencia).

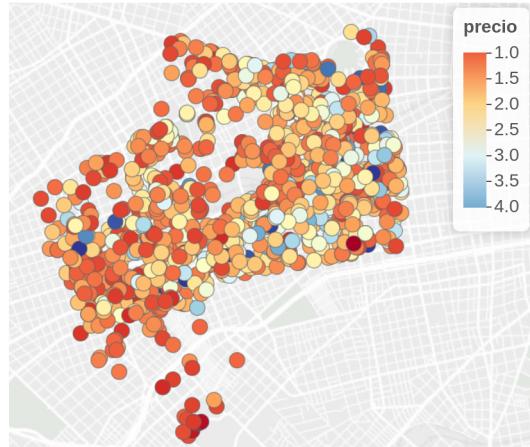
Finalmente veamos la estimación de la componente determinística del modelo, la tendencia lineal espacial del proceso.



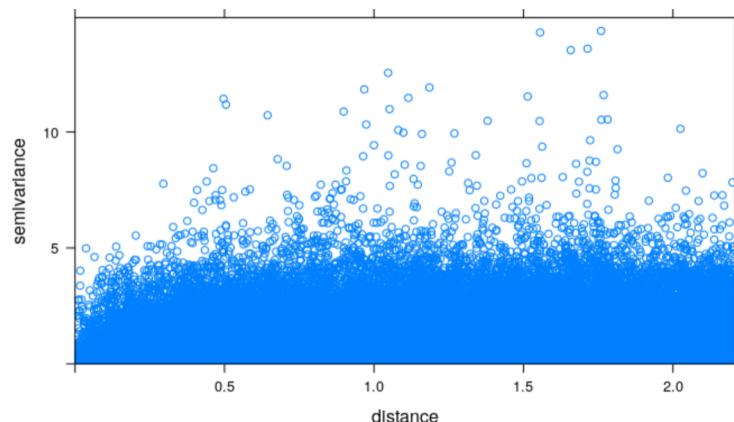
Se aprecia un gradiente más pronunciado en la componente x_1 comparado con la componente x_2 .

Ejemplo de Kriging a Datos Reales

Aplicaremos ahora la técnica de kriging a unos 1378 departamentos que se encuentran a la venta en la base de Properati, para los barrios de Flores y Caballito. El mapa que sigue muestra la ubicación de los departamentos, junto con sus precios en unidades de u\$s 100.000 (para simplificar la notación).

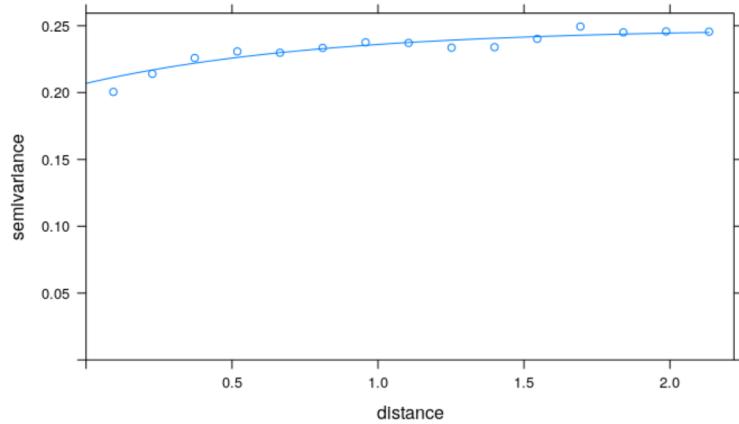


El variograma tipo “cloud” se presenta en el próximo gráfico.



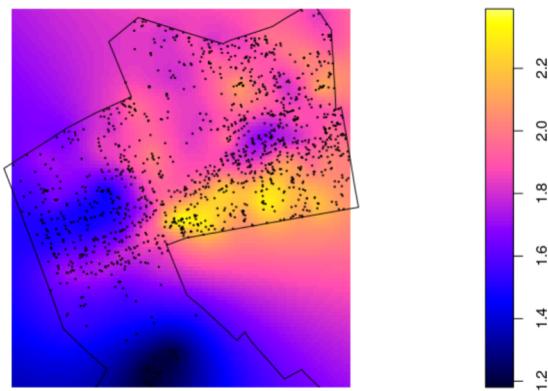
Se aprecia que existe una alta variabilidad para las muy pequeñas distancias, pese a que el análisis se realizó con los residuos de un modelo que contempla la inclusión de la covariable superficie cubierta y una componente espacial lineal. Este hecho parece razonable si se considera que estamos dejando fuera una gran cantidad de factores explicativos, que podrían reducir la variabilidad si se los incluyiera en el modelo.

A continuación ajustamos un modelo de variograma exponencial a los datos



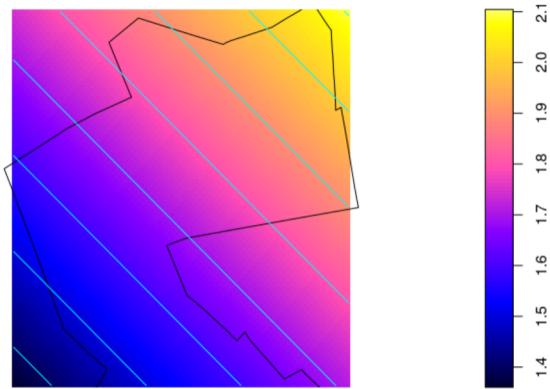
Como mencionamos previamente, la alta variabilidad en distancias cortas se modela en este caso con un nugget relativamente alto.

Así, el modelo que estimaremos contempla la componente espacial lineal, la superficie del departamento como covariable predictora, y la estructura de autocorrelación entre errores siguiendo un modelo exponencial. El mapa de predicción de kriging, para un departamento de 70 metros cuadrados, puede verse en la próxima figura



Es evidente que el máximo de predicción se dá en el suroeste de la región de análisis. Las predicciones varían en más de u\$s 100.000 entre las zonas más caras y las más baratas.

La tendencia espacial estimada se muestra en el próximo gráfico



El gradiente de precios apunta al noreste de forma suave, con una variabilidad espacial de unos u\$s 60.000 en toda la extensión de la región de estudio.

Para mejorar la comprensión de la interpolación, suele ser útil superponer a la superficie de predicción un mapa geográfico de la zona.



En este caso se ve claramente el incremento del valor predicho (unos u\$s 70.000) de un departamento de 70 metros cuadrados, cuando nos desplazamos apenas 1 kilómetro de Flores a Caballito.

Geographically Weighted Regression (GWR)

GWR es una técnica de regresión que modela en forma paramétrica la relación entre el target Y y los features X a través de parámetros que dependen potencialmente del espacio (con coordenadas u y v). En fórmula

$$Y_i = \boldsymbol{\beta}^T(u_i, v_i) \mathbf{X}_i(u_i, v_i) + \epsilon_i$$

El exceso de notación en la escritura del vector de predictoras $\mathbf{X}_i(u_i, v_i)$ responde a la necesidad de enfatizar que cada atributo (o marca) de un elemento de la muestra tiene una ubicación específica en el espacio (u_i, v_i) .

La flexibilidad agregada en esta técnica al modelo de regresión tradicional es la de permitir que los parámetros puedan variar en función del espacio, simultáneamente permitiendo que la relación determinística entre la Y y las X s también cambie en el espacio. Por supuesto, no todas las covariables X s deben tener relaciones con dependencia espacial. De hecho, el modelo lineal tradicional, en el que los coeficientes no dependen del espacio es un caso particular de este modelo.

Claramente, es imposible estimar un conjunto de parámetros distinto para punto muestral del espacio. Esto llevaría a tener muchos más parámetros que datos observados. Por lo que algún tipo de restricción a la libertad de los parámetros será necesaria. La heurística usual en problemas espaciales (y temporales) es la de pensar que ámbitos espacialmente cercanos comparten comportamientos similares. Es así que la estimación de los coeficientes se realiza mediante ventanas móviles en las que, centradas en cada punto de observación, se define un sistema de ponderadores. En cada punto las observaciones más cercanas reciben una mayor ponderación que las más alejadas. La técnica de ventanas también puede definirse usando entornos definidos por los k vecinos más cercanos, creando de esta forma ventanas adaptativas que dependerán de la densidad espacial de observaciones en cada punto del espacio.

Sea cual sea la estrategia de elección de ponderadores del entorno, la estimación se reduce simplemente a un problema de mínimos cuadrados pesados, en donde la matriz de pesos dependerá del punto (u, v) de análisis. Así la estimación de los coeficientes puede hallarse mediante la ecuación

$$\hat{\boldsymbol{\beta}}(u, v) = (X^T W(u, v) X)^{-1} X^T W(u, v) \mathbf{Y}$$

La elección de la matriz W suele obedecer a una lógica basada en las distancias, de modo tal que $W(u, v)$ es una matriz diagonal en la que los elementos de la diagonal $\{w_{i,i}\}$ cumplen con $w_{i,i} = g(d((u, v), (u_i, v_i)))$, en la que (u_i, v_i) es la ubicación de la observación i -ésima, y $g : \mathbb{R} \rightarrow \mathbb{R}$ es alguna función decreciente (o no creciente) arbitrariamente definida.

Así definida, es claro que la matriz $W(u, v)$ cambia para cada punto del espacio, pues cambian todas las distancias de (u, v) a todos los puntos de la muestra ($d((u, v), (u_i, v_i))$). Por este motivo, el estimador GWR puede pensarse como un estimador de mínimos cuadrados pesados, **PERO** los pesos cambian para cada punto.

Una vez obtenida la estimación de los parámetros para un punto del espacio $\hat{\beta}(u, v)$, es trivial calcular la predicción para ese mismo punto

$$\hat{Y}(u, v) = \hat{\beta}(u, v)^T \mathbf{X}(u, v)$$

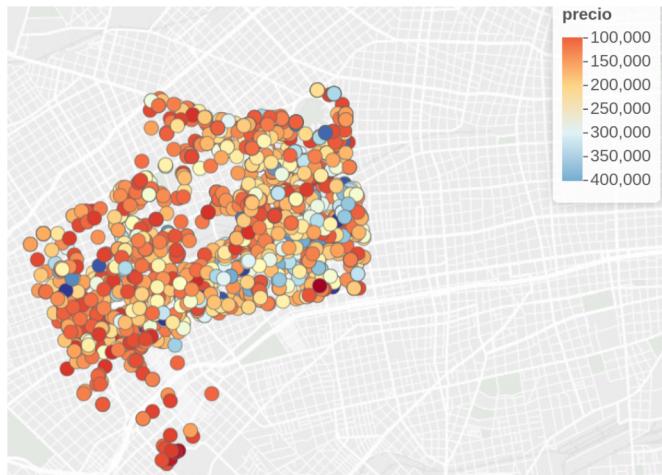
Tipicamente, no poseemos mediciones de atributos en puntos genéricos del espacio, por lo que el predictor suele evaluarse en las ubicaciones de los puntos de la muestra, es decir

$$\hat{Y}(u_i, v_i) = \hat{\beta}(u_i, v_i)^T \mathbf{X}(u_i, v_i)$$

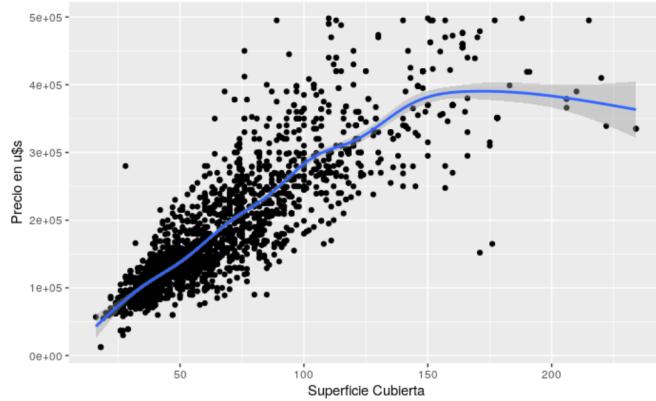
No debe sorprendernos saber que este mismo mecanismo de estimación proveé naturalmente de estimaciones de errores de predicción, y de los parámetros estimados (veasé Fotheringham (2002)).

Ejemplo: Modelado del Precio de los Departamentos

Contamos con la información (precios de venta, superficie cubierta, cantidad de cuartos, etc.) de 2248 departamentos en venta en los barrios de Flores y Caballito de la Ciudad Autónoma de Buenos Aires. La ubicación de las propiedades se muestra en el siguiente gráfico, con los colores de los puntos indicando el precio



Puede verse claramente que los precios se desparraman por los barrios, en parte por las diferencias de características de los mismos, en especial por la superficie cubierta. De hecho, un gráfico de dispersión del precio en función de la superficie muestra una relación lineal fuerte.

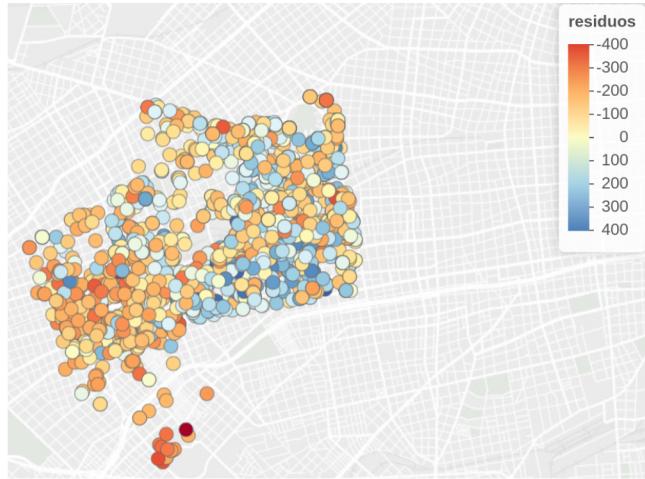


Parece razonable entonces proponer un modelo lineal que explique el precio de los departamentos en función de la superficie cubierta.

$$Precio_i = \mu + \beta Superficie_i + \epsilon_i$$

El modelo produce un buen ajuste, con un valor de $R^2 = 0.75$.

Sin embargo, como es de esperar, este ajuste global no da cuenta de la componente espacial que seguramente está presente en este fenómeno. En efecto, si graficamos los residuos (escalados por la raíz cuadrada) del modelo en el mapa obtenemos lo siguiente

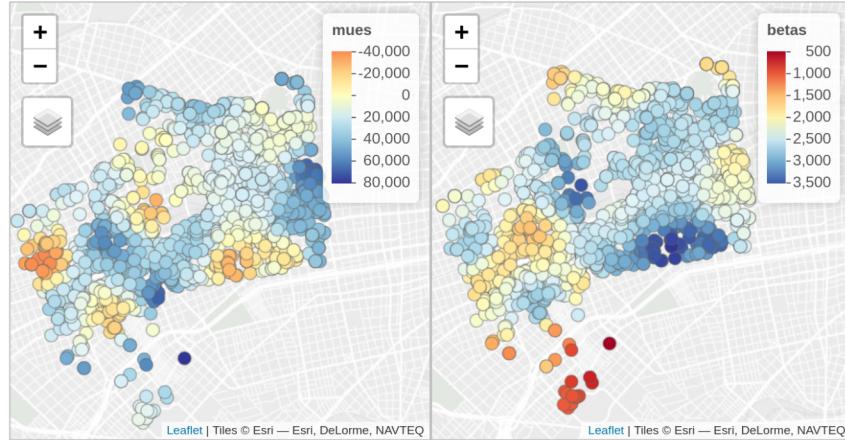


Nótese como existe un comportamiento espacial sistemático de los residuos. Los puntos se hallan fuertemente agrupados según su color (valor), en especial en la zona sur del mapa. Esto significa que el modelo global, al no incorporar el componente espacial, está perdiendo efectividad en la estimación, al no incorporar el aspecto espacial.

Apliquemos ahora la técnica de GWR, usando el mismo modelo del precio como función de la superficie cubierta. En este caso, el modelo podría escribirse como

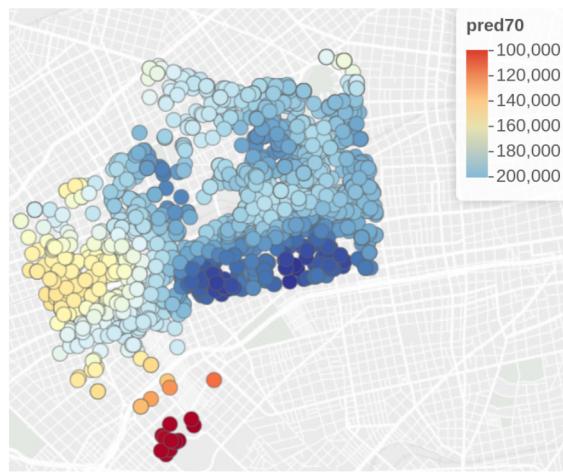
$$Precio_i = \mu(u_i, v_i) + \beta(u_i, v_i) Superficie_i + \epsilon_i$$

Los coeficientes estimados (localmente) se presentan en el próximo gráfico



No sólo ambos coeficientes varían espacialmente, sino que ambos lo hacen en forma opuesta, produciendo rectas de regresión distintas para cada zona o lugar.

Finalmente, podemos producir mapas de precios predichos para departamentos con la misma superficie,



Es evidente que la componente espacial juega un rol fundamental en la valuación de los departamentos, incluso cuando controlamos (dejamos fija) la característica más importante, la superficie cubierta.

Spatial Autoregression Models (SAR)

Estos modelos plantean explicitamente la autocorrelación existente entre las observaciones de la variable Y , proponiendo una “inercia” espacial que depende de la cercanía entre las observaciones W , a saber

$$Y = \mu 1 + \beta X + \rho W(Y - \mu 1) + \epsilon$$

donde

Este modelo puede estimarse en R mediante la función “lagsarlm”