Introducción a la Estadística Espacial

Dr. Andrés Farall

December 27, 2021

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, generalemente sucede que su comportamiento (distribución D) depende fuertemente del componente espacial S, es decir $D\left(Y/S\right) \neq \mathrm{D}(Y)$.
- 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, generalemente sucede que su comportamiento depende de Y_2 con localización espacial S_2, es decir $D\left(Y_1/X, S, Y_2\right) \neq \mathrm{D}(Y_1/X, S)$, .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

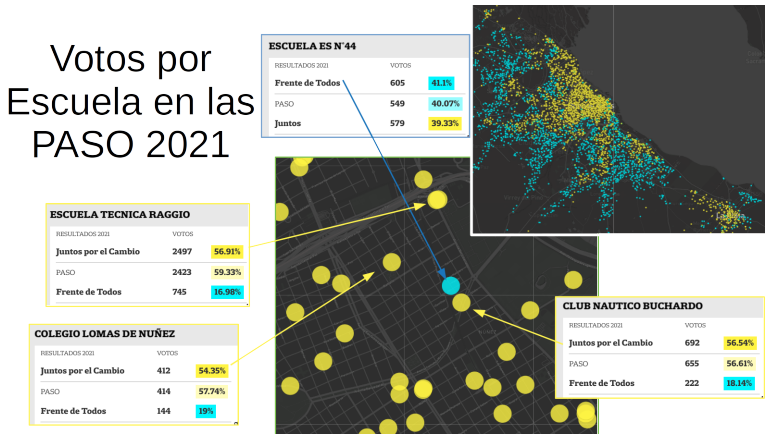
$$\left\{Z\left(s\right) : s \in S \subseteq R^d\right\}$$

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 $Z\left(s_1\right), Z\left(s_2\right) \ldots Z\left(s_n\right)$, es decir conocer

$$P\left(Z\left(s_1\right) \in A_1, Z\left(s_2\right) \in A_2 \ldots Z\left(s_n\right) \in A_n\right)$$

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 conourbano 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 cuales 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 $\boldsymbol{Z}$ 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 ocurrión en la posicón s y $Z(s) = 0$ si no hay evento en la posición s. Pese a esto, a este caso tan particular se lo conoce

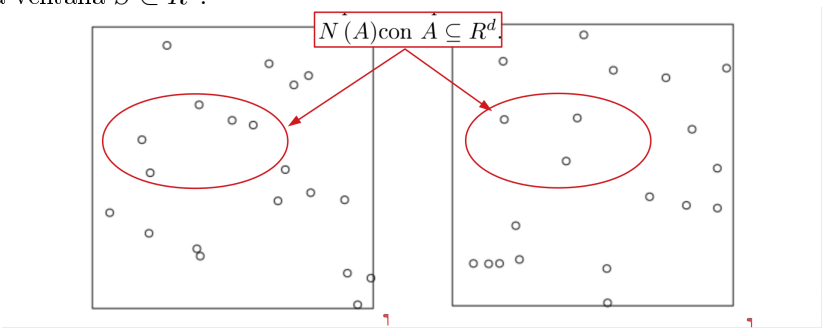
como Procesos Espaciales Puntuales (Point Pattern), y se los trabaja de manera totalmente diferente, como veremos a continuación.

Procesos Espaciales Puntuales

Un proceso estocástico puntual $\boldsymbol{X}$ es un proceso que genera conjuntos finitos aleatorios de puntos en R^d. Una definición más formal de un proceso estocástico puntual $\boldsymbol{X}$ 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 $A \subseteq R^d$.

Sorprendentemnete, la distribución del proceso $\boldsymbol{X}$ 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 ?

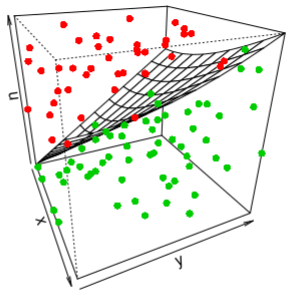
Muy facil. Si la región S es rectangular $S = [a,b] \times [c,d]$, basta con generar n pares de observaciones uniformes en los intervalos $[a,b]$ y $[c,d]$ de manera independiente.

¿ 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):

$$\text{Si } (\boldsymbol{X}, U) \sim Unif\left((\boldsymbol{x}, u) : 0 \leq u \leq f(\boldsymbol{x})\right) \text{ entonces } \boldsymbol{X} \sim f$$

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 $\sim Unif(\boldsymbol{x}, u)$ en el cubo, y nos quedamos con (aceptamos) las observaciones que satisfacen $0 \leq$

$u \leq f(\boldsymbol{x})$, 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 un un soporte más irregular que un rectangulo 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 rectangulo 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 porceso binomial, esa función es la densidad f del proceso. 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

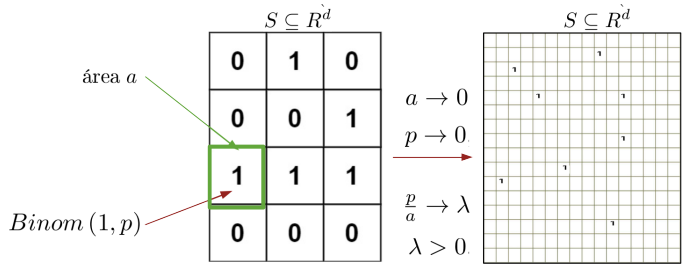
Pensemos ahora en un proceso similar al binomial, pero reemplazando la función de densidad por una función de intensidad λ, para el cual le pedimos que para toda región compacta $A \subseteq S \subseteq R^d$ cumpla que:

- $N(A) \sim P\left(\lambda = \int_A \lambda(x)\,dx\right)$. O sea que la cantidad de eventos debe comportarse como una poisson.
- Dado $N(A) = n$, $N(B) \sim Binom\left(n, \int_B \lambda(x)\,dx\right)$ 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 binomal.

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

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

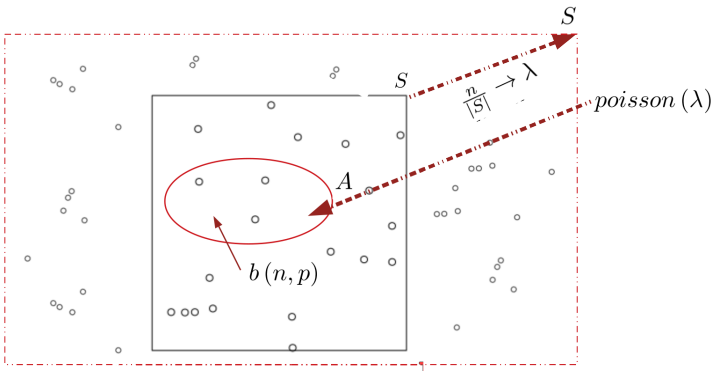
Al proceso de poisson puede llegarse de múltiples maneras, una modalidad particularmente intuitiva (al menos para mí), es partir de un proceso discreto en un soporte acotado $S \subseteq R^d$. Armemos una grilla regular con celdas de área a. En cada celda aleatorizamos la ocurrencia del evento mediante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



En el límite de este ejercico obtendremos un proceso de Poisson homogeneo 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 $\frac{n}{|S|} \to \lambda$ constante (la intensidad). La variable aleatoria $N(A)$ tiene ahora una distribución $poisson\,(\lambda)$.



¿ 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 $N \sim P\left(\lambda = \int_A \lambda(x)\,dx\right)$ como la cantidad de eventos en la región.
- Simulamos n (realización de la v.a. N) eventos según un proceso binomial con densidad $f(x) = \frac{\lambda(x)}{|S|}$

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

A modo de ejemplo mostramos en el próximo gráfico dos realizaciones de un proceso de poisson inhomogeneo. 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

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

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

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

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:

- Genrar eventos "padres" provenientes de un proceso de poisson homogeneo con función de intensidad λ.
- Para cada evento "padre" $x \in R^d$ generado en el paso anterior, generar una cantidad poisoon 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".

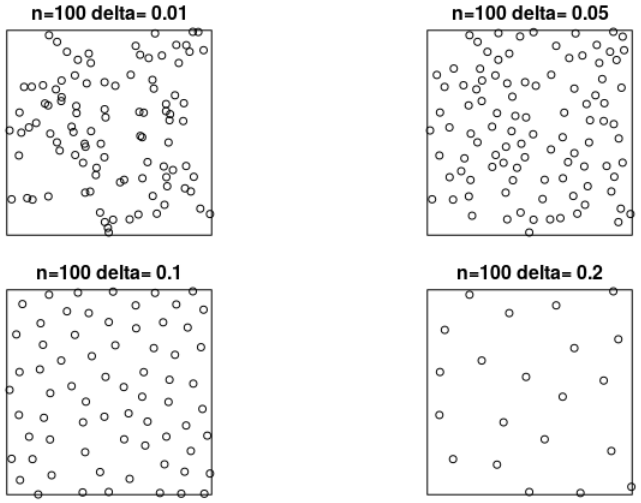
Loa parámetros que definen este proceso son: la función de intensidad λ,la cantidad esperada de hijos μ y la dispersión σ.

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 $\boldsymbol{S} \subseteq \boldsymbol{R^d}$ donce 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.

Complete Spatial Randomnes (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.

Los procesos 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 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.

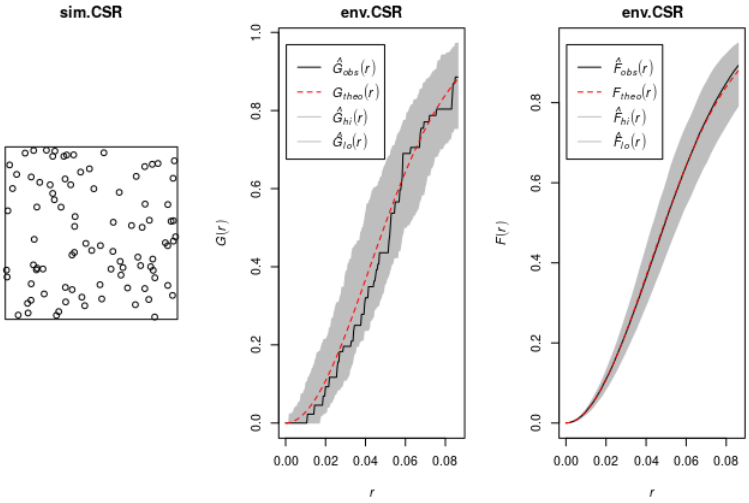
La Función F

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 vacias 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

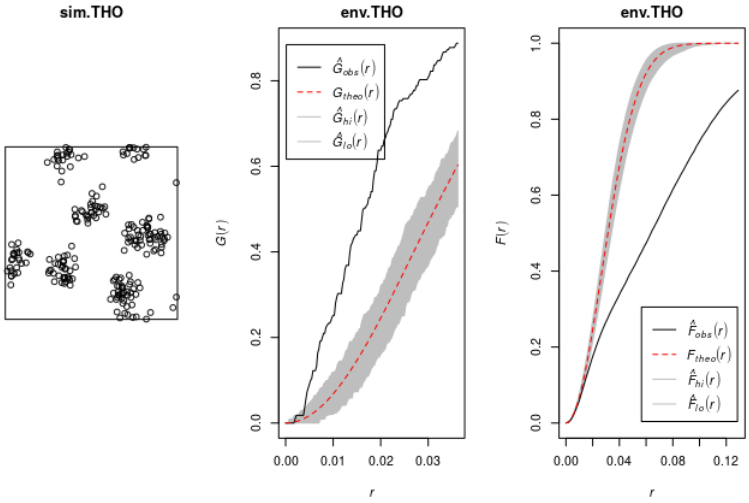
la función de distribución de la variable (aleatoria) "distancia al evento más cercano" de todos los puntos de la región de análisis.

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

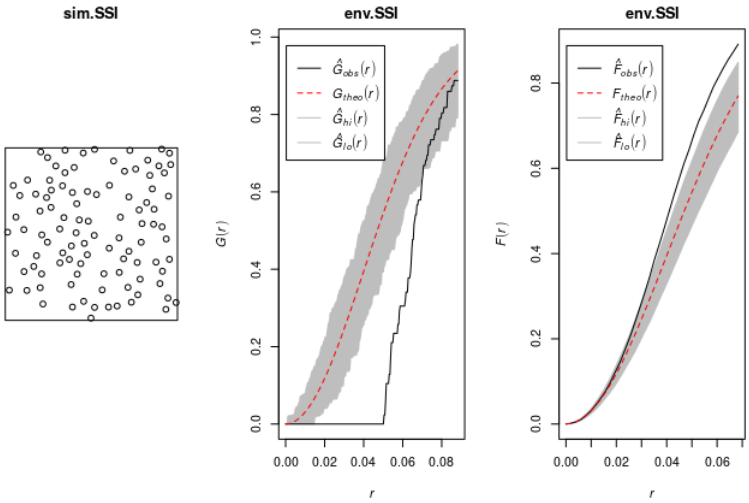
Para el proceso de Poisson homogeneo obtenemos:



Para el proceso de Thomas obtenemos:



Para el proceso de Inhibición Simple obtenemos:



Nótese que la función F en mucho mas suave que la función G, y mantiene una relación inversa con respecto al benchmark de CSR.

La Función K de Ripley

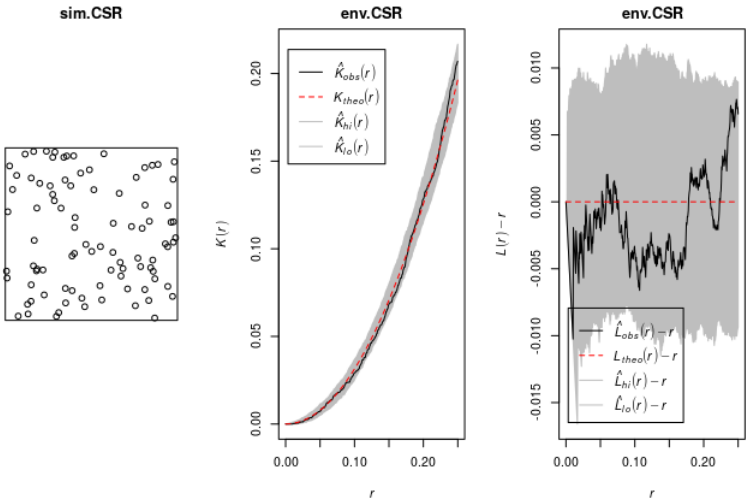
La métrica más difundida para cuantiicar el alejamiento de un patrón de eventos en relación a un proceso CSR es la función K de Ripley. Esta función $K(r)$ mide, salvo una constante, la proporción de distancias entre eventos d_{ij} que son menores a r. O sea

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

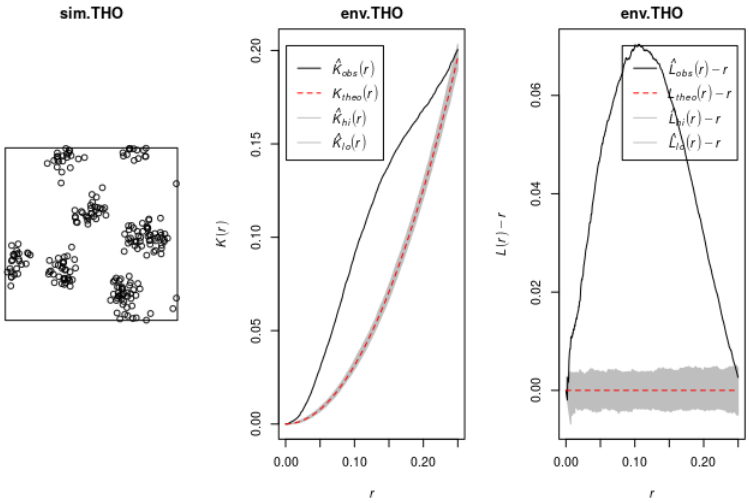
Para un proceso CSR, la función equivale a $K(r) = \pi r^2$. Para simplificar la comparación, se aplica una simple transformación a esta función, obteniendose la función $L(r) = \sqrt{\frac{K(r)}{\pi}}$. Con esta última función se suele graficar $r - L(r)$ versus r, y buscar alejamientos del valor 0.

Nuevamente, observemos para los procesos definidos previamente el comportamiento

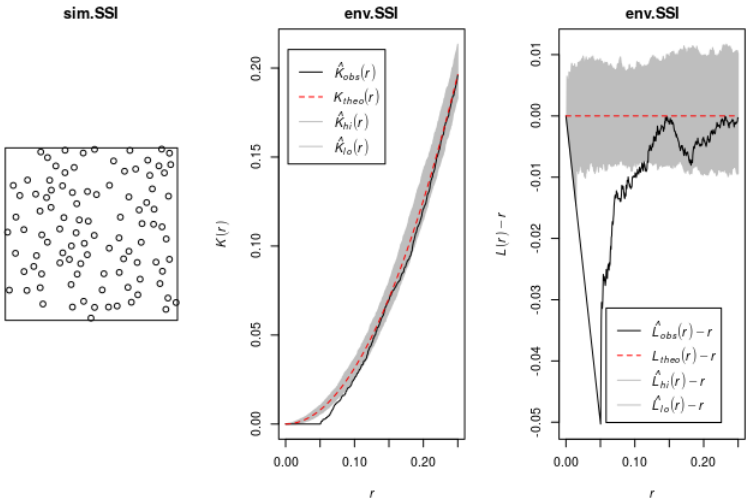
Para el proceso de Poisson homogeneo obtenemos:



Para el proceso de Thomas obtenemos:



Para el proceso de Inhibición Simple obtenemos:


sim.SSI
env.SSI
env.SSI
$\hat{K}_{obs}(r)$
$K_{theo}(r)$
$\hat{K}_{hi}(r)$
$\hat{K}_{lo}(r)$
$K(r)$
0.00
0.05
0.10
0.15
0.20
0.00
0.10
0.20
r
$\hat{L}_{obs}(r)-r$
$L_{theo}(r)-r$
$\hat{L}_{hi}(r)-r$
$\hat{L}_{lo}(r)-r$
$L(r)-r$
0.01
0.00
-0.01
-0.02
-0.03
-0.04
-0.05
0.00
0.10
0.20
r

Nótese que

Estimación de la Intensidad del Proceso

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

La técnica más difundida de estimación de la intensidad es la de KDE (Kernel Density Estimation). El estimador se define como

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

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 estima la intensidad.

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

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

Este estimador de la intensidad, es equivalente el estimador de desnsidad 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).

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 homogeneo.

El Índice de Moran

Mide la outocorrelación espacial de una variable. Se define

$$I = \frac{n}{\sum_{i,j} W_{ij}} \frac{\sum_{i,j} W_{ij}\left(X_i - \overline{X}\right)\left(X_j - \overline{X}\right)}{\sum_i \left(X_i - \overline{X}\right)}$$

El Variograma

Sea un proceso espacial $Z(s)$. Tomemos dos elementos del proceso, ubicados en s_1 y en s_2. El variograma mide la dependencia espacial entre estos elementos, de la siguiente manera:

$$\begin{aligned} \gamma(s_1, s_2) &= Var\left(Z(s_1) - Z(s_2)\right) = \\ &= E\left(Z(s_1) + E\left(Z(s_1)\right) - Z(s_2) - E\left(Z(s_1)\right)\right)^2 \end{aligned}$$

Si el proceso es homgeneo entonces $E(Z(s_1)) = E(Z(s_2))$ por lo que

$$\gamma(s_1, s_2) = E(Z(s_1) - Z(s_2))^2$$

Si el proceso es estacionario e isotrópico entonces podemos eliminar los índices y pensar que $Z(s_1) = Z(s)$ y $Z(s_2) = Z(s+h)$, por lo que

$$\gamma(s_1, s_2) = E(Z(s) - Z(s+h))^2 = \gamma(h)$$

Es claro que $\gamma(0) = 0$ y $\gamma(h) \geq 0 \forall h$.

Por supuesto que si contamos con una sóla realización del proceso, la expresión anterior no será de mucha utilidad, pues sólo contamos con finitos valores de h para calcular esta expresión, así ue tendremos que definir un estimador empírico del variograma, o suponer algín modelo paramétrico para el mismo.