Temas para la Educación
revista digital para profesionales de la enseñanza
Nº 5 - Noviembre 2009 | Federación de Enseñanza de CC.OO. de Andalucía | ISSN: 1989-4023 | Dep.Leg.: GR 2786-2008

Agregaciones espaciales.

Introducción

La complejidad de los modelos medioambientales es que virtualmente cualquier proceso físico lleva consigo una variabilidad en el espacio y tiempo.

En este artículo consideraremos solamente la **componente espacial**, cuyos métodos geoestadísticos son métodos descriptivos en el sentido que no hay ninguna interpretación causativa asociada. Ya han sido estudiados estos modelos desde la componente temporal y desde ambas componentes combinadas en anteriores artículos de los números 2, 3 y 4.

¿En qué contextos son útiles estos modelos?

La interacción supone que los casos cercanos están en el espacio. Este tipo de patrones es muy útil cuando se desea investigar **enfermedades transmisibles**. También se pone de manifiesto la interacción cuando la causa de la enfermedad es la exposición a un agente geográficamente localizado (una sustancia tóxica, ciertos tipos de radiaciones, etc.).

Una de las **necesidades** que plantea el comportamiento de estas entidades es el perfeccionamiento de los sistemas de vigilancia, de manera que puedan identificar cuando una agregación de casos de enfermos observada en un área geográfica determinada, en un período de tiempo limitado, o teniendo en cuenta ambos escenarios a la vez, es superior a lo esperado, y si ello representa un brote.

La Epidemia de cólera en Londres en 1854 fue estudiada por Dr. John Snow que realizó este mapa y que supuso el primer estudio epidemiológico. Los puntos muestran los casos de muerte. Las cruces representan los pozos de agua de los que bebieron los enfermos.

Teniendo en cuenta la gran cantidad de datos y el gran número de contrastes de hipótesis a realizar se utilizará, como soporte informático, el software libre Epidat 3.1. Este podrá ser descargado en la dirección de internet http://dxsp.sergas.es , aunque viene adjunto en la publicación.

Los métodos reseñados serán para la detección de agregaciones espaciales:

- Método Grimson
- Método Ohno
- Kriging

Para detectar agregaciones espaciales utilizan como criterio las “distancias”, de modo que objetos próximos constituyen el mismo grupo y distancias grandes separan grupos.

El investigador dispone de dos tipos diferentes de datos: tasas correspondientes a áreas de variado tamaño (barrios, por ejemplo) o coordenadas de localización geográfica, las cuales se obtienen, por ejemplo, a partir de la dirección de residencia de los casos. En dependencia del tipo de dato disponible se explora cierto patrón espacial. Cuando se dispone de tasas se trata de determinar si las áreas con altas tasas forman clusters espaciales; si se dispone de coordenadas se busca determinar cuándo se agrupan localizaciones específicas de casos. Los dos métodos incluidos en Epidat para la detección de agregaciones espaciales, los de Grimson y Ohno, utilizan tasas de áreas geográficas.

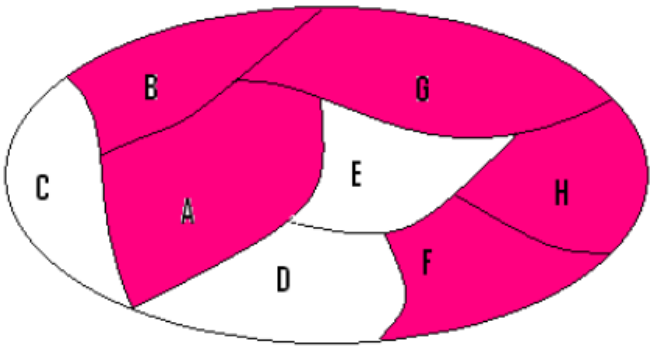
1. Método Grimson

Esta técnica ha sido diseñada para la detección de agregaciones espaciales de alguna enfermedad, dentro de una región subdividida en áreas más pequeñas, cuando las áreas de mayor riesgo tienden a agruparse espacialmente.

La prueba compara el número de pares de áreas adyacentes de alto riesgo con un número esperado, o valor crítico, bajo el supuesto de que dichas áreas de alto riesgo estuvieran distribuidas aleatoriamente en la región de estudio.

Supongamos, por ejemplo, que se conocen las tasas de incidencia de una enfermedad en cada uno de los municipios de cierta provincia, con la distribución reflejada en la Figura 1.

Figura 1.- División en municipios de la provincia de estudio.



Las áreas marcadas son las consideradas de alto riesgo de la enfermedad, cuestión que se decide definiendo un punto de corte en las tasas de incidencia de la enfermedad, y que se supone que es la tasa de incidencia más pequeña que se considera alta. Por tanto, los municipios A, B, F, G y H son aquellos con tasas de incidencia iguales o mayores que el valor tomado como punto de corte. Se observa, además, que son adyacentes A con B, A con G, B con G, G con H y H con F; es decir, existen cinco pares de áreas adyacentes de alto riesgo en la provincia. La cuestión estriba, entonces, en decidir si este hecho se da por azar o se debe a que realmente existe una agregación espacial de estas áreas.

En resumen, la **hipótesis nula** de esta prueba establece que las áreas de alto riesgo están distribuidas aleatoriamente dentro de la región de estudio, y esta hipótesis se rechaza si el número observado de pares de áreas adyacentes de alto riesgo es mayor de lo esperado.

Los **datos** que se necesitan para la aplicación del método Grimson son:

- La identificación de cada una de las áreas (suelen utilizarse letras mayúsculas).
- Los índices que identifican, para cada área, las áreas que son adyacentes a ella, es decir, con las que comparte fronteras.
- Las tasas de incidencia de cada área.
- El punto de corte en las tasas de incidencia que define las áreas de alto riesgo.

Con estos datos se realizan los siguientes **cálculos**:

- N: el número total de áreas.
- n: el número de áreas de alto riesgo.
- P: el número de pares de áreas adyacentes de alto riesgo.
- ni: el número de áreas adyacentes al área i (i=1, ..., N) y la media y varianza de estos N valores:

$$\mu = \frac{\sum_{i=1}^{N} n_i}{N} ; \sigma^2 = \frac{\sum_{i=1}^{N} (n_i - \mu)^2}{N}$$

En esta situación, la probabilidad de observar P o más áreas adyacentes de alto riesgo, es decir, el p-valor de la prueba, se obtiene a partir de la distribución de Poisson. Cuando n es pequeño comparativamente con N, se puede admitir que P sigue una **distribución de Poisson**, cuestión que puede verificarse comprobando si la media y la varianza de P son aproximadamente iguales, E(P)=Var(P).

Expliquemos las características básicas de esta importante distribución.

Siméon Denis Poisson (1781-1840), fue un físico y matemático francés al que se le conoce por sus diferentes trabajos en el campo de la electricidad, también hizo publicaciones sobre la geometría diferencial y la teoría de probabilidades.

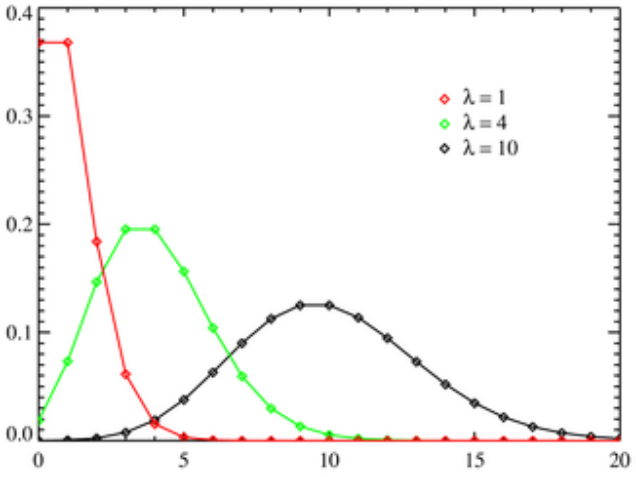
En 1837 publicó en su trabajo *Recherches sur la probabilité des jugements en matières criminelles et matière civile* ("*Investigación sobre la probabilidad de los juicios en materias criminales y civiles*"). en el cual describe la probabilidad como un acontecimiento fortuito ocurrido en un tiempo o intervalo de espacio bajo las condiciones que la probabilidad de un acontecimiento ocurre es muy pequeña, pero el número de intentos es muy grande, entonces el evento ocurre algunas veces.

El trabajo estaba enfocado en ciertas variables aleatorias N que cuentan, entre otras cosas, un número de ocurrencias discretas (muchas veces llamadas "arribos") que tienen lugar durante un intervalo de tiempo de duración determinada. Si el número esperado de ocurrencias en este intervalo es λ, entonces la probabilidad de que haya exactamente *k* ocurrencias (siendo *k* un entero no negativo, *k* = 0, 1, 2, ...) es igual a:

$$f(k;\lambda) = \frac{e^{-\lambda}\lambda^k}{k!},$$

dónde

- *e* es el base del logaritmo natural (*e* = 2.71828...),
- *k*! es el factorial de *k*,
- *k* es el número de ocurrencias de un evento,
- λ es un número real positivo, equivalente al número esperado de ocurrencias durante un intervalo dado. Por ejemplo, si los eventos ocurren de media cada 4 minutos, y se está interesado en el número de eventos ocurriendo en un intervalo de 10 minutos, se usaría como modelo una distribución de Poisson con λ = 2.5.



Su media (esperanza) y su varianza son:

$$\mu = \lambda$$

$$\sigma^2 = \lambda$$

La distribución de Poisson puede ser vista como un caso limitante de la distribución binomial. Cuando λ tiende a infinito, podemos aproximar a una distribución normal. Por ello, podemos tipificar ya que conocemos cual es la media y varianza de una Poisson.

La distribución de Poisson, se aplica a varios fenómenos discretos de la naturaleza (esto es, aquellos fenómenos que ocurren 0, 1, 2, 3, ... veces durante un periodo definido de tiempo o en un área determinada) cuando la probabilidad de ocurrencia del fenómeno es constante en el tiempo o el espacio. Ejemplos de estos eventos que pueden ser modelados por la distribución de Poisson incluyen:

- El número de autos que pasan a través de un cierto punto en una ruta (suficientemente distantes de los semáforos) durante un periodo de tiempo.
- El número de errores de ortografía que uno comete al escribir una única página.
- El número de llamadas telefónicas en una central telefónica por minuto.
- El número de servidores web accedidos por minuto.
- El número de animales muertos encontrados por unidad de longitud de ruta.
- El número de mutaciones de determinada cadena de ADN después de cierta cantidad de radiación.
- El número de núcleos atómicos inestables que decayeron en un determinado periodo de tiempo en una porción de sustancia radiactiva. La radiactividad de la sustancia se debilitará con el tiempo, por lo tanto el tiempo total del intervalo usado en el modelo debe ser significativamente menor que la vida media de la sustancia.
- El número de estrellas en un determinado volumen de espacio.
- La distribución de receptores visuales en la retina del ojo humano.
- La inventiva de un inventor a través de su carrera.

Dichos valores pueden calcularse de forma exacta a partir de m y s, según el método propuesto por Grimson. En esta situación, la probabilidad de observar P o más áreas adyacentes de alto riesgo, es decir, el **p-valor** de la prueba, se obtiene a partir de la distribución de Poisson.

$$Valor\ p = \Pr[Poisson(E(P)) \geq P]$$

A medida que n aumenta, se hace mayor la diferencia entre los valores E(P) y Var(P) y, en ese caso, puede utilizarse la distribución normal para calcular el p-valor:

$$Valor\ p = \Pr\left[N(0,1) \geq \frac{P - E(P)}{\sqrt{V(P)}}\right]$$

Las **limitaciones** del método Grimson son:

- Esta técnica otorga el mismo peso a todas las áreas que comparten fronteras con un área determinada, sin tener en cuenta la longitud de estos bordes. Esto podría introducir un sesgo en el análisis, por el hecho de que las áreas con fronteras más extensas tienen mayor probabilidad de conformar agregaciones que otras áreas con bordes más cortos. Este efecto podría neutralizarse, en alguna medida, tomando áreas lo más homogéneas posibles dentro de la región de estudio.
- La sensibilidad del método depende, en gran medida, de la elección del punto de corte para distinguir las áreas de alto riesgo.

Ejercicio: En una región con 11 áreas sanitarias se observaron, durante el año 2000, las tasas de mortalidad por cáncer de colon que se presentan. Si una tasa mayor o igual a 14 por 100.000 define un área de alto riesgo ¿existe evidencia de alguna agregación espacial entre esas áreas?

Los datos pueden introducirse desde un fichero en formato Dbase, Access o Excel, con una configuración como la mostrada en la tabla.

AREA	INDICES*	TASAS
A	B,G	8,0
B	A,C,D,F,G	9,5
C	B,D,K	15,7
D	B,C,E,F	21,1
E	D,F,J,K	8,9
F	B,D,E,G,I,J,K	4,4
G	A,B,F,H,I	7,7
H	G,I	12,1
I	F,G,H,J,K	8,0
J	E,F,I	2,1
K	I,F,E,C	18,9

* Áreas adyacentes al área correspondiente

En el ejemplo, el punto de corte se ha tomado a 14, por lo que se tienen tres áreas de alto riesgo (C, D y K) cuyas tasas son superiores a dicho valor. Así mismo, se tendrán sólo dos pares de áreas adyacentes de alto riesgo (CK, CD), ya que entre D y K no hay contigüidad. Los resultados con Epidat 3.1 nos dan:

```
Vigilancia: Detección de clusters, agregaciones espaciales

Archivo de trabajo: C:\Archivos de programa \Epidat 3.1 \Ejemplos
\Vigilancia \Grimson.xls

Campo que contiene:
  Área: AREA
  Índices de áreas adyacentes: INDICES
  Tasas: TASAS
Número de áreas     : 11

Método Grimson
Pares de áreas adyacentes de alto riesgo
----------------------------------------
Observadas:       2
Esperadas :  1,2000
Probabilidad { >= 2} =  0,1683
```

Los resultados indican que la probabilidad de observar 2 ó más pares de áreas adyacentes de alto riesgo es superior al valor convencional de 0,05, por tanto, no hay evidencia para rechazar la hipótesis nula a un nivel de significación del 5%. Se concluye, entonces, que las agregaciones observadas se han producido por mero azar.

2. Método Ohno.

Este método ha sido desarrollado26 para identificar patrones geográficos de mortalidad o morbilidad observados visualmente en el mapa de una región dividida en áreas más pequeñas (municipios, comarcas, o áreas sanitarias, por ejemplo).

Si las áreas adyacentes tienden a presentar niveles de la enfermedad más similares de lo que se esperaría por azar, entonces existe evidencia de agregación espacial.

Para la aplicación de la técnica de Ohno, se necesitan los siguientes **datos**:

- La identificación de cada una de las N áreas en que se divide la región de estudio (suelen utilizarse letras mayúsculas).
- Los índices que identifican, para cada área, las áreas que son adyacentes a ella, es decir, con las que comparte fronteras.
- Las tasas de incidencia de cada área.
- El número de categorías (k) en las que se quieren agrupar las áreas en función de sus tasas de incidencia, y los k-1 puntos de corte de las tasas de incidencia que definen esas categorías. La decisión de qué valores tomar compete al investigador o grupo de investigadores, basándose en la experiencia y conocimiento de la enfermedad objeto de estudio, y debe abarcar una gama de riesgos, desde los menores hasta los más elevados observados en la región de estudio.

El primer paso del método consiste en clasificar las áreas según el nivel de riesgo de enfermedad que presenta cada una, es decir, atendiendo a los puntos de corte definidos para las tasas de incidencia. Así, a partir de los k-1 puntos de corte Vi (i=1,...,k-1) y las tasas de incidencia de la enfermedad para cada área Ti (i= 1,...,N), las k categorías de riesgo quedan establecidas de la siguiente tabla:

Categoría	Punto de corte	Definición	Número de áreas
1	V_1	$T_i<V_1$	N_1
2	V_1, V_2	$V_1\leq T_i<V_2$	N_2
...	...	...	...
k-1	V_{k-2}, V_{k-1}	$V_{k-2}\leq T_i<V_{k-1}$	N_{k-1}
k	V_{k-1}	$T_i\geq V_{k-1}$	N_k

Luego se calcula el número de pares de áreas, dentro de cada categoría, que son adyacentes, es decir, el número de pares de áreas que cumplen a la vez dos condiciones:

- Están en el mismo nivel de riesgo (son *concordantes*)
- Tienen contigüidad espacial (son *adyacentes*).

El método de Ohno consiste en comparar, mediante un estadístico Ji-cuadrado, el número observado de pares de áreas *adyacentes y concordantes* en la categoría i-ésima (ACi, i=1, ..., k) con el número esperado (EACi, i=1, ..., k) bajo la hipótesis de que la distribución es uniforme en toda la región de estudio:

$$\chi^2 = \left(\frac{AC_i - EAC_i}{\sqrt{EAC_i}} \right)^2 , \text{i=1, ..., k}$$

donde:

$$EAC_i = \frac{A}{\frac{N(N-1)}{2}} \frac{N_i(N_i-1)}{2} , \text{i=1, ..., k}$$

y A el número total de pares de áreas adyacentes.

En esta última fórmula el primer factor representa la proporción de pares de áreas en toda la región de estudio que son adyacentes, y el segundo es el número máximo de pares de áreas concordantes (de las cuales no necesariamente todas son adyacentes) que pueden formarse en la categoría i-ésima.

También se realiza una prueba global para comparar el número observado de áreas adyacentes y concordantes de toda la región (AC) con su número esperado (EAC), que de nuevo se calcula bajo la hipótesis de que dichas áreas se distribuyen uniformemente en la región:

$$\chi^2 = \left(\frac{AC - EAC}{\sqrt{EAC}} \right)^2$$

donde:

$$AC = \sum_{i=1}^{k} AC_i \text{ y } EAC = \sum_{i=1}^{k} EAC_i$$

El objetivo de esta prueba global es la detección de agregaciones espaciales en toda la región, sin considerar una categoría de riesgo específica.

Las **limitaciones** del método Ohno son:

- Al igual que el método de Grimson, esta técnica otorga el mismo peso a todas las áreas que comparten fronteras con un área determinada, sin tener en cuenta la magnitud de esas fronteras.
- La aproximación Ji-cuadrado del estadístico es válida si el número de pares de áreas adyacentes es mucho más pequeño que el número total de pares posibles: N(N-1)/2.

Ejercicio: En una región con 10 áreas de salud se observaron, durante el año 2000, las tasas de mortalidad por cáncer de colon que se presentan en la Tabla 9; los datos se encuentran también en el archivo OHNO.xls de los ejemplos de Epidat 3.1. El equipo de investigadores ha decidido utilizar cuatro categorías de nivel de riesgo de la enfermedad, para lo cual definen tres puntos de corte en las tasas: 2, 6 y 10, ¿existe evidencia de alguna agregación espacial entre esas áreas?

Los datos pueden introducirse desde el teclado (entrada manual) o importarse en formato Dbase, Excel o Access (entrada automática) con una configuración mostrada en la tabla.

AREA	INDICES	TASAS
A	B,G,H	7
B	A,G,F,D,C	8,5
C	B,D	11,9
D	C,B,F,E	15,7
E	D,F,J	1,5
F	B,G,I,J,E,D	4,4
G	B,A,H,I,F	7,7
H	A,G,I	1,1
I	H,G,F,J	8,3
J	I,F,E	3,3

En cualquiera de los dos casos debe introducirse, además, el número de categorías de riesgo y los puntos de corte en las tasas de incidencia para definir esas categorías.LOResultados con Epidat 3.1:

```
Vigilancia: Detección de clusters, agregaciones espaciales

Archivo de trabajo: C:\Archivos de programa \Epidat 3.1 \Ejemplos
\Vigilancia \OHNO.xls

Campo que contiene:
  Área: AREA
  Índices de áreas adyacentes: INDICES
  Tasas: TASAS

Número de áreas     : 10
Nº de categorías    : 4
Método Ohno

                Áreas adyacentes concordantes
Categoría   Áreas   Observadas    Esperadas Ji-cuadrado  Valor p
---------   -----  ------------ ------------ -----------  -------
        1       2            0       0,4222      0,4222   0,5158
        2       2            1       0,4222      0,7906   0,3739
        3       4            4       2,5333      0,8491   0,3568
        4       2            1       0,4222      0,7906   0,3739
---------   -----  ------------ ------------ -----------  -------
    Total      10            6       3,8000      1,2737   0,2591
```

En los resultados, Epidat presenta las categorías de riesgo con el número de áreas incluidas en cada una de ellas. También incluye el número observado y esperado de pares de áreas adyacentes y concordantes para cada nivel de riesgo y, por último, los valores $\chi 2$ y los valores p correspondientes a las pruebas parciales y global. En el ejemplo, no se observa evidencia de agregación espacial en ninguna categoría ni globalmente.

3.- Kriging.

El **krigeaje** o **krigeado** (del francés *krigeage*) es un método geoestadístico de estimación de puntos que utiliza un modelo de variograma (El **variograma** o **semivariograma** es una herramienta que permite analizar el comportamiento espacial de una variable sobre un área definida, obteniendo como resultado la influencia de los datos a diferentes distancias. para la obtención de datos). Calcula los pesos que se darán a cada punto de referencias usadas en la valoración. Esta técnica de interpolación se basa en la premisa de que la variación espacial continúa con el mismo patrón. Fue desarrollada inicialmente por Danie G. Krige a partir del análisis de regresión entre muestras y bloques de mena, las cuales fijaron la base de la geoestadística lineal.

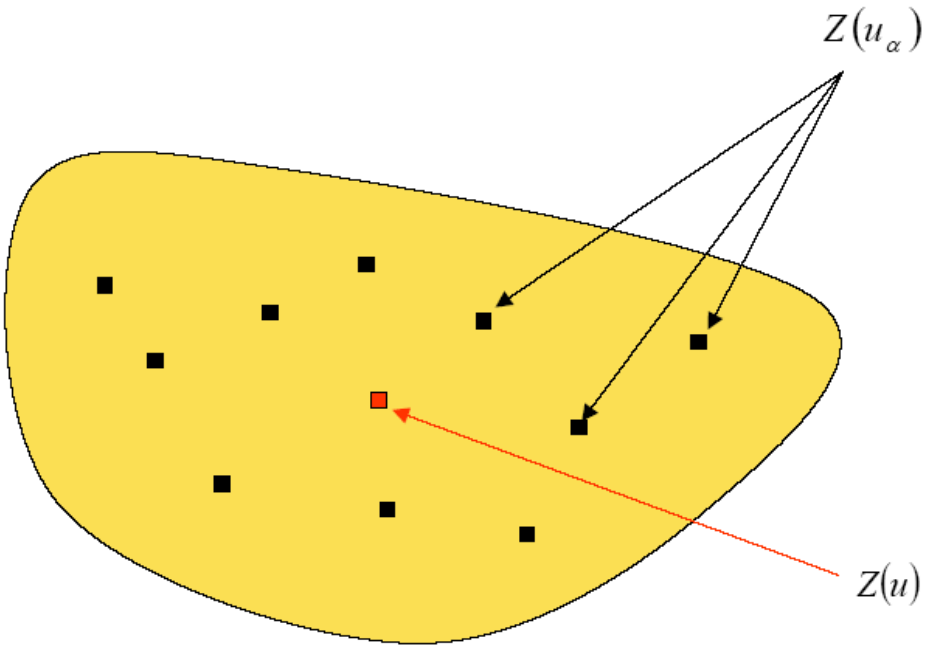
Danie G. Krige (Daniel Gerhardus Krige) nacido en el Estado Libre de Orange en Sudáfrica, es un ingeniero de minas pionero en el campo de la geoestadística y fue profesor en la University of the Witwatersrand en Sudáfrica. Los trabajos empíricos de Krige para evaluar recursos mineros fueron formalizados en la década de 1960 por el ingeniero francés Georges Matheron.

Matheron desarrolló la técnica denominada kriging basada en la labor investigadora previa de Krige en el campo de la geoestadística.

Llamamos Kriging a las técnicas de predicción lineal para datos geoestadísticos (entre otros). La situación que tenemos es la siguiente:

$$Z : \mathbb{R}^d \to \mathbb{R}$$
$$s \in D \subseteq \mathbb{R}^d$$

Consideremos $Z(u_\alpha), \alpha = 1,2,\dots N$ puntos en los cuales se tiene información de determinada propiedad en el yacimiento y $Z^*(u)$ la estimación de $Z(u)$ a partir de los puntos $Z(u_\alpha)$



Planteamiento básico de la estimación por Kriging:

Considerar la estimación de $Z(u)$ como una combinación lineal de las observaciones disponibles

$$Z^*(u) = \sum_{\alpha=1}^{N} \lambda_\alpha(u) Z(u_\alpha)$$

y escoger los pesos bajo un criterio en el cual se considera que dicha estimación es óptima. Este es que el estimador sea insesgado y que

$$\mathrm{var}\left[Z(u) - Z^*(u)\right]$$ sea mínima

Bibliografía

- Webs:

http://www.wikipedia.org
http://www.google.com

- Libros y artículos de consulta:

Raubertas RF. *"Spatial and temporal analysis of disease occurrence for detection of clustering. Biometrics".* 1988;44:1121-9.

CLUSTER 3.1 Software System for Epidemiologic Analysis. Instruction Manual. February 1993. U.S. Department of Health & Human Services. Public Health Service. Agency for Toxic Substances and Disease Registry. Atlanta, Georgia.

Barton DE, David RN. *"The random intersection of two graphs".* In: David FN, editor. Research papers in Statistics. New York: Wiley; 1966. p. 445-59.

Aldrich TE. *"Detecting space-time aggregation of rare events".* Am J Epidemiol. 1984;120:464.

Vigilancia en Salud Pública. Epidat 3.1.