Imagen cortesía de la NASA

MÉTODOS ESTADÍSTICOS AVANZADOS USADOS EN LOS Sistemas de Información Geográfica

(I. Métodos no paramétricos)

Esperanza Ayuga (2008)

Índice

- Introducción
- Métodos no paramétricos
- Procesos estocásticos
- Modelo Lineal General

Introducción

La tecnología de los Sistemas de Información Geográfica (SIG) y de la Teledetección son, por su propia naturaleza, campos multidisciplinares.

Cada disciplina ha desarrollado la terminología y metodología que reflejan el interés concreto de su campo. Las herramientas de análisis espacial son, por tanto, de una diversidad notable.

Introducción

Las herramientas de análisis estadístico espacial abarcan un amplio campo de métodos orientados a resolver problemas espaciales diferentes:

- realce de una imagen,
- reconocimiento de patrones,
- interpolación de datos para localizar depósitos minerales,
- la investigación espacio-temporal de la evolución de una enfermedad,
- la modelización de tendencias socioeconómicas relacionadas con las migraciones,

Introducción

Las herramientas estadísticas se caracterizan por:

- primar la realidad en la representación espacial mediante los SIG, por encima de la sencillez de las técnicas;

- ser de amplia aplicación en todas las disciplinas;

- ser computacionalmente factibles y emplear, principalmente, la exploración y estimación de la heterogeneidad espacial.

Introducción

Estas técnicas se clasifican según las características estructurales de los datos y la dimensionalidad del estudio a realizar (considerando una o más variables espaciales)

Introducción

<table>
<tr><th rowspan="2">Características estructurales de los datos</th><th colspan="2">Dimensionalidad</th></tr>
<tr><th>Univariante</th><th>Multivariante</th></tr>
<tr><td rowspan="2">De situación</td><td>Método de puntos próximos
Funciones-k</td><td>Funciones k-bivariantes
Interacciones espacio-tiempo</td></tr>
<tr><td colspan="2">Estimación núcleo de la densidad
Regresión núcleo
Alisado bayesiano</td></tr>
<tr><td rowspan="3">De valor</td><td>Autocorrelación espacial
Correlogramas
Variogramas</td><td>Correlación espacial multivariante</td></tr>
<tr><td>Krigeado</td><td>Co-krigeado</td></tr>
<tr><td>Modelo lineal general</td><td>Mod. lineal general espacial
Agrupaciones
Correlaciones canónicas</td></tr>
<tr><td>Interdependientes</td><td colspan="2">Modelos de interacción espacial</td></tr>
</table>

Introducción

Lo que llamamos datos de situación son aquellos que representan la presencia o ausencia de un suceso en una localización determinada (*procesos puntuales*).

(ESRI)

Introducción

Por ejemplo en una zona en estudio, localización de los puntos donde se da una enfermedad y en el caso multivariante, dónde hay enfermos y dónde ambulatorios. También puede considerarse una variable temporal, p. e. en el caso anterior, dónde se da la enfermedad en un grupo de años consecutivos.

Introducción

Los denominados datos de valor son variables aleatorias, cualitativas o cuantitativas, asociadas a un conjunto de localizaciones. Estas localizaciones pueden ser puntos específicos, cuadrículas o polígonos en mallas regulares o irregulares.

Introducción

Por ejemplo, variables que midan diferentes características del suelo en puntos muestreados, medidas de usos de suelo obtenidas por teledetección sobre mallas regulares o ratio de mortalidad en polígonos irregulares.

En el caso multivariante se miden varias características o variables en cada localización y una de las variables puede ser el tiempo.

Introducción

Los datos interdependientes son variables cuantitativas, tales que cada una de ellas está asociada a un “enlace” o a un par de localizaciones. Éstas suelen ser dos puntos pero también pueden considerarse conjuntos de puntos o áreas regulares o irregulares.

Introducción

Por ejemplo flujo de individuos de distintos lugares de residencia para realizar compras.

En el caso multivariante se puede asociar al punto origen del enlace un conjunto de medidas que determinen la demanda del producto que se va a comprar.

Introducción

Las herramientas estadísticas que se recogen en este tema se agrupan en:

- técnicas de estimación no paramétrica (*puntos próximos* y *métodos de estimación núcleo*);
- procesos estocásticos (*autocorrelaciones, correlogramas, variogramas, krigeado*) y
- el modelo lineal general,con aplicaciones numerosas (*modelo lineal general, agrupaciones, correlaciones canónicas, etc.*)

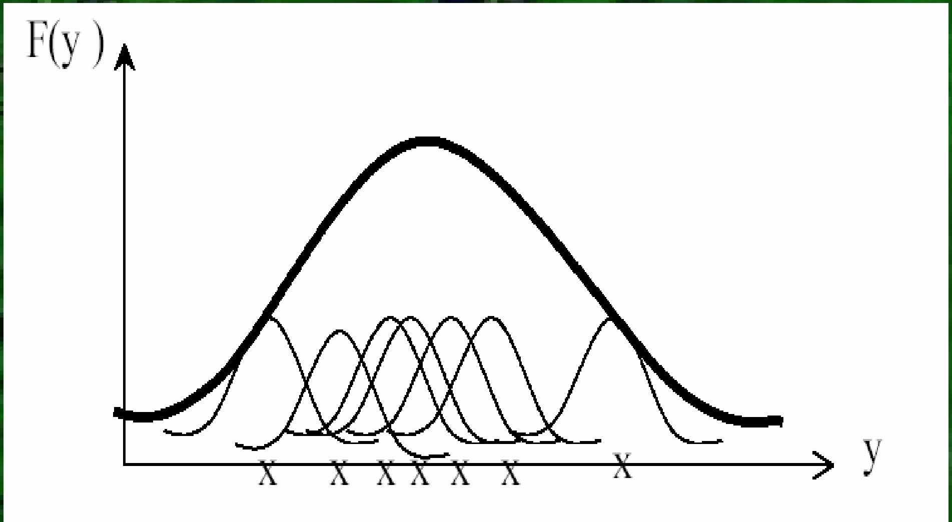
Introducción

Repasaremos estos tres bloques con sus herramientas más significativas y de mayor aplicabilidad en el campo de la Teledetección y los SIG.

Estimación no paramétrica

La estimación no paramétrica consiste en la estimación de valores poblacionales basándose en hipótesis muy generales sobre éstas.

La forma de la distribución es una de las estimaciones que se realizan.



Estimación núcleo de la función de densidad univariante

Los estudios de estimación no paramétrica de la función de densidad han sido muy numerosos desde que se demostró la convergencia casi segura del histograma de frecuencias a la densidad.

De los diferentes procedimientos de estimación, los mejores estudiados matemáticamente -y aquellos para los que existe un mayor número de aplicaciones a datos reales - son los basados en la definición de una función núcleo

Para su empleo es necesario elegir tanto el núcleo como un valor del parámetro de alisado. Ambos determinarán la expresión final de la función de densidad estimada.

Estimación núcleo

El núcleo es una función K(x), a partir de la cual se puede establecer el siguiente estimador no paramétrico de cualquier función de densidad f(x)

$$\boxed{\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)}$$

Donde h es el parámetro de alisado y $X_1, \ldots, X_n$ los datos observados.

Estimación núcleo

La función núcleo podrá ser cualquier función que cumpla ciertas propiedades (Nadaraya,1989) que son esencialmente, las mismas condiciones que cualquier función de densidad, y que garantizan unas buenas propiedades de la estimación.

- K simétrica

$$\int K(t)\,dt = 1$$

$$\int tK(t)\,dt = 0$$

$$\int t^2 K(t)\,dt = k_2 \neq 0$$

$$\int |K(t)|\,dt < \infty$$

$$|tK(t)| \to 0 \text{ si } |t| \to \infty$$

Estimación núcleo

El parámetro de alisado, también llamado ancho de banda, es un número positivo (h) que se determina, en general, minimizando algún tipo de error (el VC el más eficaz).

$$h_n \to 0 \quad y \quad nh_n \to \infty \qquad si \quad n \to \infty$$

entonces

$$\hat{f}(y) \xrightarrow{p} f(y)$$

Estimación núcleo de densidades multivariantes

Se define este estimador para el caso multivariante como una suma de "funciones núcleo" centradas en las observaciones.

Estimación multivariante

Si una muestra, dada por $\mathbf{X}_1, \ldots, \mathbf{X}_n$, es observada en un espacio d-dimensional, entonces la función de densidad subyacente se estima con

$$\boxed{\hat{f}_n(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)}$$

Donde h es el parámetro de alisado y K(**x**) es una función núcleo definida para **x** d-dimensional y que satisface la condición $\int_{R^d} K(x)dx = 1$.

Estimación multivariante

Una aproximación sencilla es cambiar inicialmente los datos por medio de una transformación lineal, obteniendo la matriz de covarianza unidad, a continuación se alisa usando una función núcleo radialmente simétrica, y finalmente, volver a retransformar los datos.

Estimación multivariante

Esto equivale a usar el estimador de la densidad

$$\hat{f}_n(x) = \frac{(\det S)^{-1/2}}{nh^d} \sum_{i=1}^{n} k\left[\frac{(x-X_i)^T S^{-1} (x-X_i)}{h^2} \right]$$

Donde k se toma como: $k(x^T x) = K(x)$

y S es la matriz de las covarianzas de los valores muestrales.

Estimación multivariante

En el estimador se emplea un sólo parámetro de alisado h, lo que implica que el núcleo centrado en cada punto muestral es del mismo orden en todas las direcciones del plano.

En algunas ocasiones, sería más apropiado emplear un vector de anchos de banda o, en caso de que la dispersión de los valores sea mucho mayor en una dirección de los ejes que en otra, una matriz de coeficientes reducidos.

Estimación multivariante

Mayor complicación en los cálculos de h

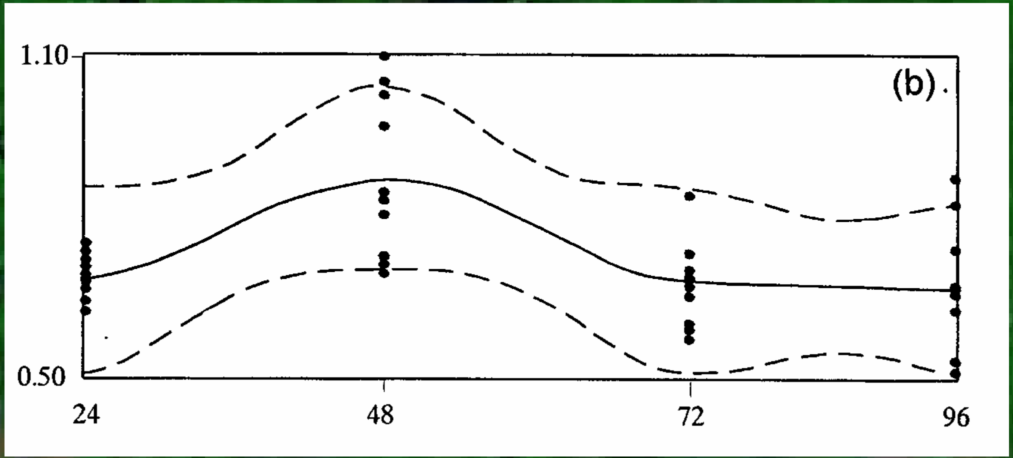
Se debe seleccionar un método de optimización para determinar h con iguales problemas que en el caso unidimensional.

Se han propuesto diferentes expresiones para las funciones núcleo multivariantes.

Estimación no paramétrica

Estimadores núcleo de la regresión.

Los estimadores núcleo para la regresión se proponen inicialmente para el modelo de regresión de diseño fijo.



Estimación no paramétrica

Asumiremos por simplicidad, que las medidas fijas X_i satisfacen:

$$\boxed{0 \leq X_1 \leq X_2 \leq \ldots \leq X_n \leq 1}$$

y que g está acotado y es derivable en [0,1]. Se proponen dos estimadores de la línea de regresión:

Estimación no paramétrica

Priestley y Chao (1972) recomiendan el estimador de expresión

$$\hat{g}_n(x, h_n) = \sum_{i=1}^{n} \frac{X_i - X_{i-1}}{h_n} K\left(\frac{x - X_i}{h_n}\right) Y_i$$

donde K es una función de densidad simétrica y de cuadrado integrable y h_n el parámetro de alisado

Estimación no paramétrica

Nadaraya (1964) y Watson (1964) introducen un estimador núcleo basado en la expresión de la esperanza condicional. El estimador determinado de esta forma es el siguiente:

$$\hat{g}_n(x, h_n) = \frac{\sum_{i=1}^{n} K\left(\frac{X - x_i}{h_n}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{X - x_i}{h_n}\right)}$$

donde h_n es el parámetro de alisado y K es una función de densidad con propiedades concretas.

Estimación no paramétrica

Método de los puntos próximos

Dado un punto **x** y fijado un entero k, sea $D_k(\mathbf{x})$ la distancia euclídea de **x** a su k-ésimo punto más próximo entre los $X_1,\ldots, X_n$, y sea $Vol_k(\mathbf{x})= C_d[D_k(\mathbf{x})]^d$ el volumen de la esfera d-dimensional de radio $D_k(\mathbf{x})$ donde C_d es el volumen de la esfera unidad d-dimensional.

Estimación no paramétrica

El estimador de densidad por puntos más próximos viene dado por

$$\hat{f}(\mathbf{x}) = \frac{k/n}{\mathrm{Vol}_k(\mathbf{x})}$$

para el caso d-dimensional, se debe elegir un k_n proporcional a $n^{4/(d+4)}$, y con constante de proporcionalidad en función de **x**.

Estimación no paramétrica

Una ventaja de éste estimador es que siempre es positivo aún en regiones donde los datos están muy dispersos.

Es adecuado para estimar la densidad en un punto pero no para la función completa de densidad ya que se comprueba que conduce a una estima de densidad discontinua y con integral infinita debido a sus grandes colas.

Referencias

- Abramson, I. S. (1984). Adaptative density flattening-A metric distortion principle for combating bias in nearest neighbor methods. *Annl. Stat*., 12, 880-886.

- Ayuga, E. (1992). Modelos no paramétricos de ajuste de curvas aplicados al ámbito forestal. Tesis Doctoral de E.T.S.I.M., U.P.M.