MÉTODOS ESTADÍSTICOS AVANZADOS USADOS EN LOS Sistemas de Información Geográfica

(II. Procesos y Modelo Lineal General)

Esperanza Ayuga (2008)

Introducción

Las técnicas estadísticas que se pueden emplear en los SIG actualmente presentan tres características fundamentales:

- su complejidad,
- su enfoque al estudio de relaciones espaciales
- su aplicabilidad a muy diferentes campos

Introducción

1. Los menos conocidos, aunque de uso creciente, son los métodos núcleo de estimación de densidades y regresión y, también, el de puntos próximos;
2. Los procesos estocásticos en el dominio del tiempo y de la frecuencia también son de recientísima aparición
3. El modelo lineal general y los métodos de análisis multivariable que se basan en él son los que más se han empleado hasta ahora.

Procesos estocásticos

En los procesos donde intervienen variables que evolucionan en el tiempo, el espacio, o en ambos a la vez (caudal de un río, cc en la atmósfera de un agente contaminante, ...)

La modelización de las observaciones requiere técnicas que consideren su dependencia.

La teoría de los procesos estocásticos es la base para el análisis de estas variables.

Procesos estocásticos

Un proceso estocástico es una colección de variables aleatorias indexadas por un conjunto T, $\{x(t)/t \in T\}$. Es habitual considerar T como un subconjunto de los números reales, bien discreto {0, 1, ... } o bien continuo [0, 4), que suele identificarse con el tiempo.

Sin embargo, nada impide considerar un conjunto de subíndices T más complejo (espacios Hilbert), como $\mathbb{R}^2$ o $\mathbb{R}^n$ lo que permitiría, una vez conocido el proceso, el estudio y predicción de procesos temporales, espaciales o en n dimensiones.

Procesos estocásticos

Las aplicaciones de los procesos estocásticos a los estudios del medio son múltiples.

1. Estudios sobre la evolución de la productividad o diversidad de un ecosistema en el tiempo
2. La evolución de parámetros de contaminación, son ejemplos de procesos estocásticos con conjunto de índices unidimensional.

Procesos estocásticos

En dos dimensiones, la elección de un conjunto apropiado de índices sirve para caracterizar distintos procesos:

Si el conjunto de índices estuviera formado por todos los posibles pares de coordenadas en el plano, para cada par de coordenadas (x, y) existiría una variable aleatoria (profundidad del suelo, precipitación...) que podría analizarse bajo la óptica de un proceso estocástico.

Procesos estocásticos

La teoría de los procesos estocásticos es, desde el punto de vista teórico, la más apropiada para tratar el tipo de información manejada por los SIG y la teledetección.

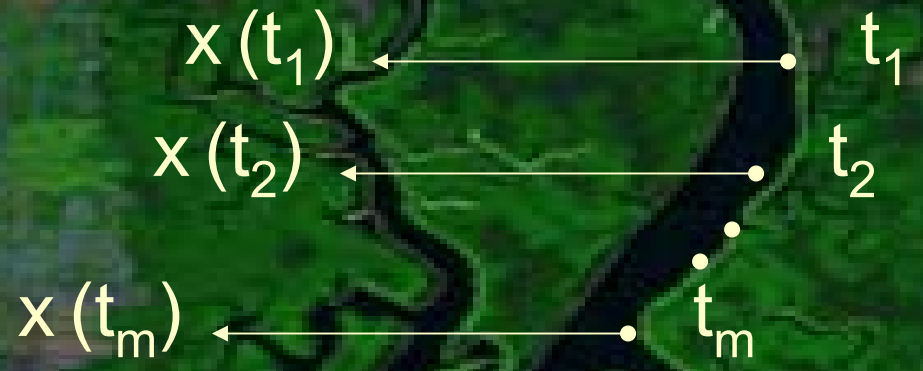
Bajo la óptica de los procesos estocásticos el concepto de muestra se complica.

Procesos estocásticos

Por una parte se puede considerar como muestra al conjunto formado por los valores que toma la variable $X(t_1)$ al repetir m veces el experimento en un punto fijo t del conjunto de índices

Procesos estocásticos

Por otra se puede considerar como el resultado de ejecutar el experimento, una sola vez, en distintos puntos del conjunto de índices (trayectoria);



el resultado de una trayectoria es la medición del experimento aleatorio en los instantes t_1, ..., t_n (o en las coordenadas X_1, ..., X_n).

Procesos estocásticos

Un elemento fundamental en el estudio de los procesos estocásticos es el estudio de la variación del proceso entre dos puntos de T. La covarianza (C) y el coeficiente de correlación (R) representan dos expresiones de esta variación,

$$C(s,t)=E\left[\ \{X(s)-E(X(s))\}\ \{X(t)-E(X(t))\}\ \right]$$

$$R(s,t)=\frac{C(s,t)}{\sqrt{C(s,s)C(t,t)}}$$

Procesos estocásticos

Los procesos gaussianos, caracterizados porque la distribución conjunta de cualquier conjunto finito de variables aleatorias

$$\{X(t_1), ..., X(t_n)\}$$

es normal multivariable, pueden describirse completamente a partir del conocimiento de la media en todo t y de la covarianza en todo par (s, t) de T.

Una propiedad importante de algunos procesos es la estacionaridad.

Procesos estocásticos

El conjunto de los números reales $\mathbb{R}$ (del cual T es un subconjunto) está siempre ordenado, y el vector que une dos puntos s y t, o su distancia, se definen unívocamente .

Por el contrario, en $\mathbb{R}^2$, o $\mathbb{R}^n$, este orden es más difícil de establecer y, en general, no existe una única definición de distancia entre dos puntos.

Procesos estocásticos

Según Chatfield (1989), pueden distinguirse dos enfoques fundamentales para el estudio de periodicidades en los procesos estocásticos:

- el análisis en el conjunto de índices (llamado también en el dominio del tiempo)
- el análisis en el dominio de la frecuencia.

Ambos análisis son complementarios y se basan en detectar pautas periódicas de variación en el conjunto de datos muestrales para extrapolarlas al resto del conjunto de índices.

Análisis en el dominio del tiempo

La especificidad de este enfoque se basa en determinar la influencia que tienen los valores que toman variables separadas k unidades en el conjunto de subíndices.

Por definición de procesos isótropos, la covarianza entre dos puntos del conjunto de índices depende sólo de la distancia entre esos puntos:

$$C(s_1, s_{1+k}) = C(s_2, s_{2+k})$$

Análisis en el dominio del tiempo

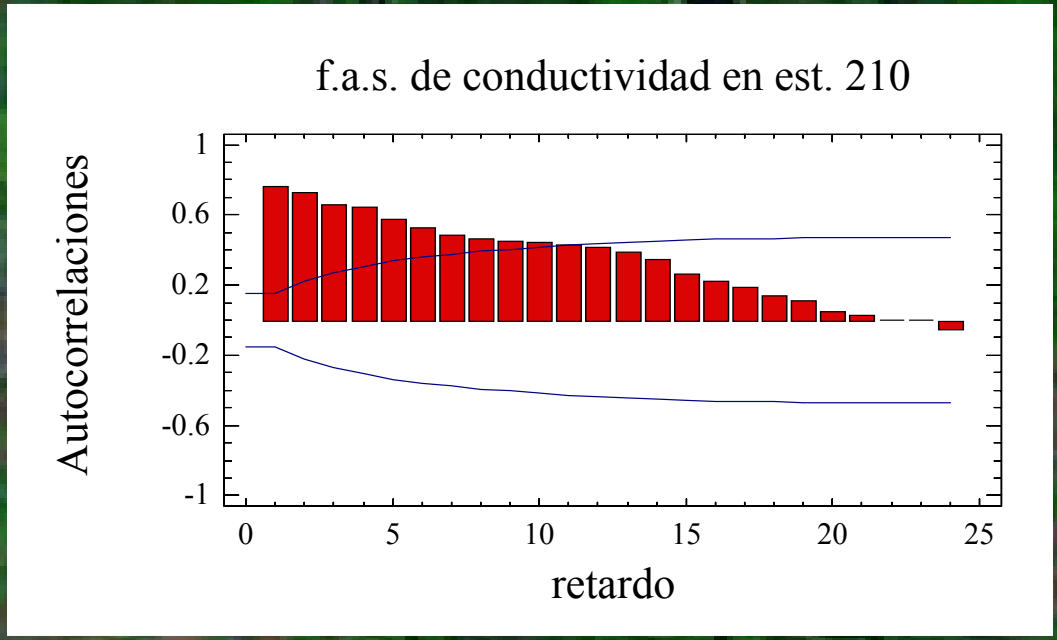
Cuando la covarianza se expresa en función de la distancia k que separa dos puntos recibe el nombre de autocovarianza, $\{\gamma(k)\}$. la distancia k de separación entre dos puntos se llama retardo.

De forma análoga a la autocovarianza se puede definir la autocorrelación entre observaciones a retardo k

$$\rho(k) = \gamma(k) / \gamma(0)$$

Análisis en el dominio del tiempo

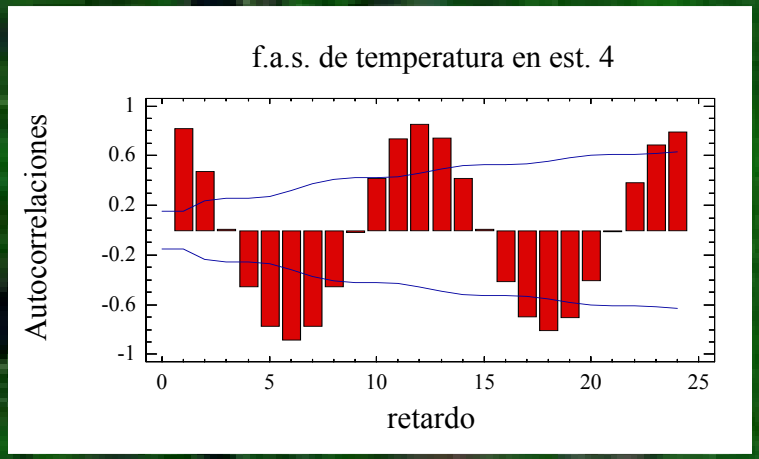
Se denomina función de autocorrelación simple (fas) o correlograma a la representación de los coeficientes de autocorrelación en función del retardo.



Análisis en el dominio del tiempo

La determinación del correlograma juega un papel destacado en la selección de un modelo adecuado para representar la serie de datos.

Además, en cualquier caso, la forma del correlograma permite detectar las variaciones periódicas en los datos a analizar.



Análisis en el dominio del tiempo

Supongamos que la información disponible viene dada por la trayectoria $\{x(t_1),..., x(t_N)\}$ procedente de medir los resultados del experimento en los instantes $\{t_1,..., t_N\}$.

A partir de estos datos se suele utilizar, para estimar la poblacional, la autocovarianza muestral:

$$c(k) = \sum_{t=1}^{N-k} \frac{(x_t - \bar{x}_N)(x_{t+k} - \bar{x}_N)}{N}$$

Análisis en el dominio del tiempo

La estimación del coeficiente de correlación r(k) se realiza por el estadístico “coeficiente de correlación muestral”.

$$r(k) = \frac{c(k)}{c(0)} = \frac{\sum_{t=1}^{N-k} (x_t - \bar{x}_N)(x_{t+k} - \bar{x}_N)}{\sum_{t=1}^{N} (x_t - \bar{x}_N)^2}$$

Los modelos para predecir la variable se pueden encontrar en el texto de Box y Jenkins (1976).

Análisis en el dominio de la frecuencia

Mientras el análisis en el dominio del tiempo se basa en la función de autocorrelación simple, el análisis en el dominio de la frecuencia se basa en el estudio de la función de densidad espectral.

El enfoque “frecuencista” considera que las variaciones no estacionarias en los procesos estocásticos se pueden simular mediante superposición de ondas sinusoidales que “vibran” con diferente frecuencia, fase y amplitud.

Análisis en el dominio de la frecuencia

El modelo expresa, mediante estas ondas, el proceso estocástico. Una forma de expresar la contribución de cada onda w_p al proceso total es considerar alguna medida relacionada con su amplitud.

Llamando R_p a la amplitud de la onda elemental w_p, ($R_p = \sqrt{a_p^2 + b_p^2}$), tenemos la medida:

$$I(w_p) = NR_p^2/4\pi$$

Análisis en el dominio de la frecuencia

Se llama periodograma a la representación de la variación de $I(w_p)$ respecto de w_p. Para una muestra de tamaño N se formula como:

$$I(w_p) = \frac{(\sum x_t \cos 2\pi pt/N)^2 + (\sum x_t \, sen 2\pi pt/N)^2}{N\pi}$$

Análisis en el dominio de la frecuencia

Desgraciadamente el periodograma no es consistente con el modelo de agregación de ondas (no sólo las w_p). Si aumenta el número de frecuencias analizadas, al incrementar infinitamente el tamaño muestral, I(w) no explica la contribución de la onda w a la variación del sistema.

El periodograma se asimila a un “sintonizador” de un receptor de radio, así, la serie que observamos sería la señal emitida por una radio y el periodograma sería el dial que busca en qué frecuencia se “oye” mejor la señal emitida.

Análisis en el dominio de la frecuencia

Una mejor aproximación se obtiene al considerar la contribución de todas las posibles ondas sinusoidales, empleando los resultados del teorema de Wienerkhintchine (Bartlett, 1955).

Este teorema relaciona la función de autocovarianza con una nueva función (distribución espectral), F(w), monótona no decreciente, que expresa la contribución aportada a la varianza del proceso por las ondas de frecuencia menor o igual a w.

Análisis en el dominio de la frecuencia

La derivada de la función de distribución espectral respecto de w se conoce como función de densidad espectral (f(w)) y está relacionada con la función de autocovarianza al ser su transformada de Fourier:

$$f(w) = \frac{1}{\pi} \sum_{-\infty}^{\infty} \gamma(k) e^{-iwt}$$

Análisis en el dominio de la frecuencia

Buenos estimadores de la densidad espectral se pueden obtener mediante transformaciones de la función de autocovarianza truncada (estimada a partir del correlograma muestral) o mediante suavizaciones del periodograma.

Dependencia en el espacio

Otra forma de modelizar procesos estocásticos con índices en $\mathbb{R}^n$, con n>1, más sencilla y más fácil de aplicar, es considerar la dependencia en el espacio desde el punto de vista de la geoestadística, que se ha visto con anterioridad.

Modelo lineal general

En sistemas que incluyen factores variables diversos resultará interesante estudiar el efecto que algunos de ellos ejercen sobre los otros.

La situación más favorable está representada por los casos en que es posible establecer una expresión funcional entre las variables del problema que interprete correctamente las relaciones existentes.

Modelo lineal general

En la mayoría de los casos reales la situación no corresponde a la descrita anteriormente. La complejidad de la dependencia entre las diversas variables implicadas en el problema resulta muchas veces demasiado grande para que pueda encontrarse una función matemática que sea su expresión; en tales casos, se busca un modelo sencillo para las dependencias existentes que sustituya aproximadamente a la función desconocida que las relaciona

Modelo lineal general

La construcción del modelo, que servirá para predecir resultados, será valiosa aún en aquellos casos en que no se vea una relación física inmediata entre las variables que figuran en el modelo, pues la ecuación matemática en que se expresa puede ser útil para predecir valores de algunas de las variables.

Modelo lineal general

El modelo que se emplea con más frecuencia es el que construimos aplicando la regresión mínimo cuadrática, que da lugar a modelos lineales con dos o más variables, aunque se consideren a veces regresiones no lineales.

Modelo lineal general

Aunque la teoría que soporta los modelos lineales en los parámetros puede aplicarse a un tipo mucho más amplio de problemas, se puede aceptar que un vector aleatorio cumple el modelo lineal cuando satisface una ecuación que incluye variables aleatorias, variables matemáticas y parámetros; y que es lineal en los parámetros y las variables aleatorias.

Modelo lineal general

Estudiamos vectores aleatorios **X** cuya distribución satisface la siguiente expresión:

$$E\left[\overline{X}\right] = \overline{\beta} A'$$

Donde $\overline{X} = (X_1, ..., X_n)$ es un vector aleatorio, $\overline{\beta}$ es un vector de parámetros desconocidos (b_1, b_2, ..., b_n) y A es una matriz nxk ($k \neq n$) de constantes a_{ij} conocidas.

Modelo lineal general

Para un análisis posterior es conveniente formular el modelo como:

$$E\left[\overline{X}\right] = \overline{\beta} A' + \overline{\varepsilon}$$

donde $\overline{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ es un vector de variables aleatorias tales que $E(\varepsilon_j)=0$.

Modelo lineal general

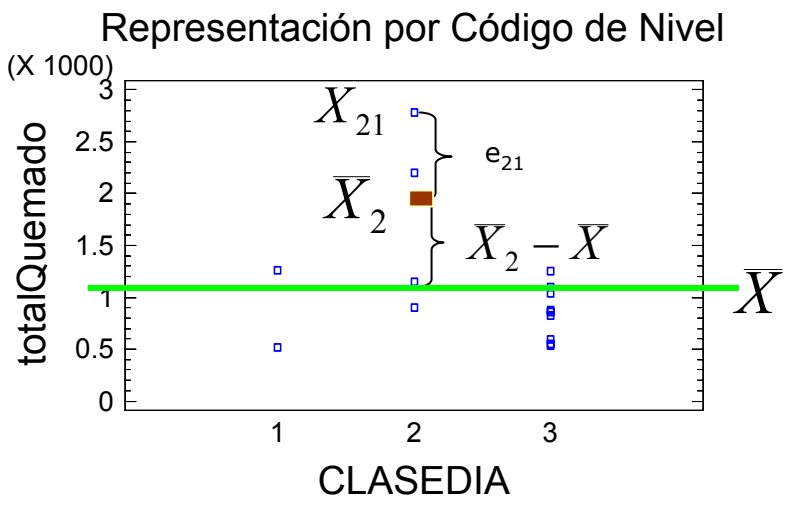
La inferencia paramétrica sobre este tipo de modelos, ha de referirse a los parámetros desconocidos. Una hipótesis sobre estos parámetros es lineal cuando puede formularse como :

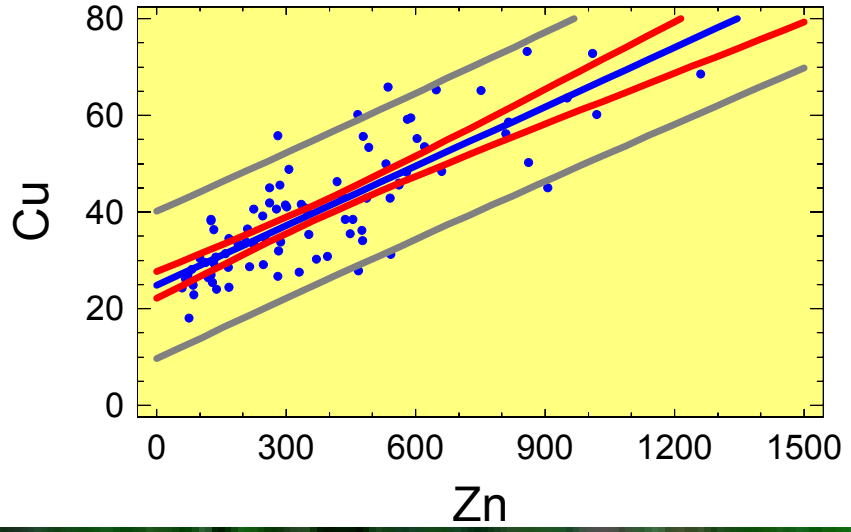
$$H_0 : \overline{\beta} H' = \overline{A}$$

Donde H' es una matriz de coeficientes conocidos nxk (k≠n) y $\overline{A}$ un vector de constantes.

Modelo lineal general

Casos especiales del contraste de hipótesis lineales son el modelo lineal de regresión y el análisis de la varianza.





Modelo lineal general

Prácticamente todos los métodos que se derivan de la teoría de la regresión lineal y múltiple junto con los conceptos de correlación simple, múltiple o parcial son los de mayor uso en el campo de la teledetección y de los SIG.

Modelo lineal general

El problema de la clasificación de un determinado individuo o punto o región del espacio en una de varias posibles categorías o conjuntos sobre la base de una serie de medidas obtenidas de esa unidad es un problema interesante que resuelve la técnica del análisis discriminante y tiene gran aplicación en los problemas de estudio de clasificación ecológica.

Modelo lineal general

Otro de los modelos lineales más interesantes es el análisis factorial, cuya finalidad es explicar un conjunto de variables observables mediante un número más reducido de variables hipotéticas llamadas factores.

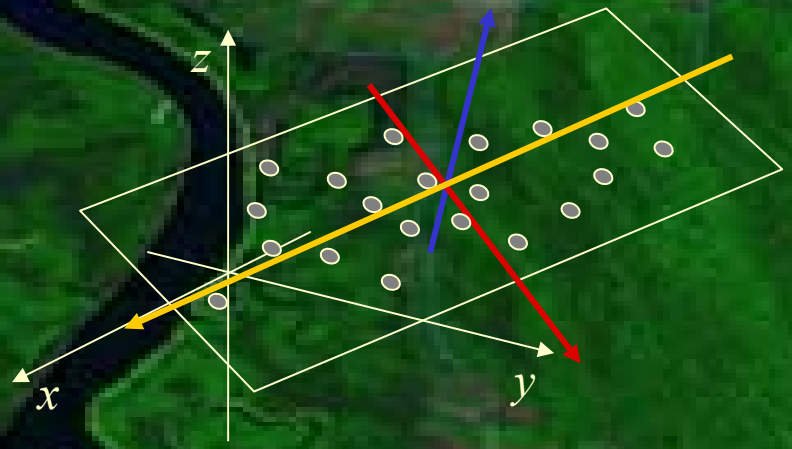
En general, los factores no se pueden observar directamente y responden a conceptos de naturaleza más abstracta que las variables originales.

Modelo lineal general

Los componentes principales, en estrecha relación con el análisis factorial, se definen como combinaciones lineales normalizadas de variables estadísticas que gozan de propiedades especiales relativas a las varianzas, de forma que esas combinaciones lineales normalizadas tengan varianzas máxima o mínima, problema equivalente a obtener los vectores propios de la matriz de covarianzas

Modelo lineal general

Desde el punto de vista estadístico, el conjunto formado por los componentes principales suministra un sistema conveniente de coordenadas, y sus respectivas varianzas caracterizan sus propiedades estadísticas.



Modelo lineal general

En la práctica estadística, el método de los componentes principales se utiliza para encontrar las combinaciones lineales que explican el máximo de variabilidad.

Para disminuir el número de variables se realiza un cambio de variable adecuado y luego se eliminan las combinaciones lineales con varianzas pequeñas considerando sólo aquellas que tengan grandes varianzas.

Modelo lineal general

El análisis de CP por su propiedad relativa a las varianzas nos permite saber qué variables, bien aisladas, bien como combinación lineal, debemos considerar en el estudio, descartando aquellas otras cuya contribución a la varianza se revela más escasa.

Modelo lineal general

Las correlaciones canónicas representan, en cambio, la dependencia que existe entre dos conjuntos de variables, por medio del estudio de las correlaciones en ambos conjuntos.

Modelo lineal general

El método determina en primer lugar las combinaciones lineales en cada conjunto que tienen correlaciones máximas; estas combinaciones lineales serán las primeras coordenadas de una nueva referencia.

Seguiremos, buscando dos nuevas combinaciones, una en cada conjunto, de tal modo que la correlación entre ambas sea máxima y además sean incorreladas con las determinadas anteriormente...

Modelo lineal general

Continuamos el proceso hasta que los nuevos sistemas estén completamente determinados. El trabajo resulta más simple si sólo hay que considerar unas pocas combinaciones lineales de las variables en cada conjunto, y deseamos que éstas sean las más fuertemente correlacionadas.

Referencias

- Peña, D. (2002). Análisis de datos multivariantes, McGraw-Hill, Madrid.

- Peña, D. (2005) Análisis de series temporales. Alianza Editorial, Madrid.