Análisis estadístico de la variabilidad geográfica de riesgos en salud pública

Juan Ferrándiz
Universitat de València
Juan.Ferrandiz@uv.es

XI Reunión Nacional de
Mejora Genética Animal
Pamplona, 14 de junio de 2002

GUIÓN

- Motivación
- Campos aleatorios sobre continuos
- Redes fijas de localidades
- Procesos puntuales
- Perspectivas

MOTIVACIÓN

Estadística espacial

- Estadística:
  - explicación de la variabilidad observada en magnitudes aleatorias
  - Predicción de futuras observaciones
- Espacial:
  - las magnitudes aleatorias van ligadas a localizaciones geográficas
  - muy relevante en estudios medioambientales

Referencias fundamentales

- Ripley (1981) *Spatial Statistics.* Wiley
- Diggle (1983) *Statistical analysis of spatial point patterns.* Academic Press
- Cressie (1991) *Statistics for spatial data.* Wiley
- Lawson *et al.* (1999) *Disease mapping and risk assessment for Public Health.* Wiley

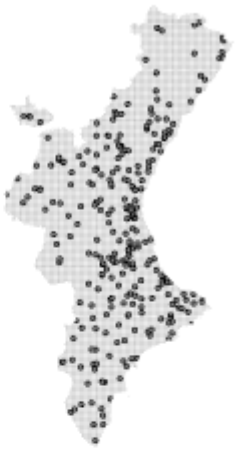
Tipos de datos

DATOS SOBRE

- Localización continua (pluviometría, polución ambiental, ...)

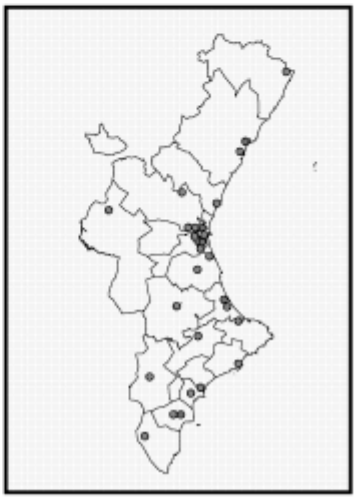
Estaciones meteorológicas

- 282 estaciones meteorológicas en la Comunidad Valenciana
- Registro diario de pluviometría, temperatura, etc.
- Estudio agregado para un periodo determinado
- Ojalá fueran también de captación de contaminación ambiental



Red de vigilancia de la gripe

- Red de médicos voluntarios para diagnóstico preciso y rápido de gripe
- Se pretende anticipar los brotes agudos
- Se puede calibrar con el sistema de EDO (Declaración Obligatoria, lento pero fiable)
  - Generalistas
  - Pediatras



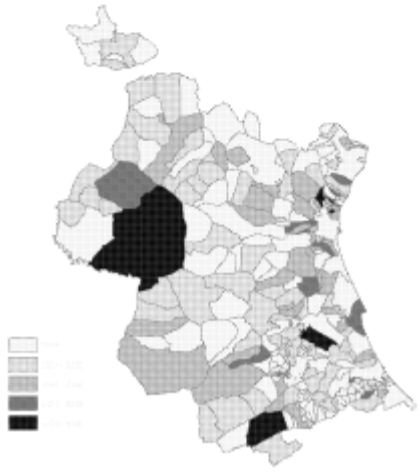
Tipos de datos

DATOS SOBRE

- Localización continua (pluviometría, polución ambiental, ...)
- Parrilla fija (estadísticas municipales, imágenes digitales,...)

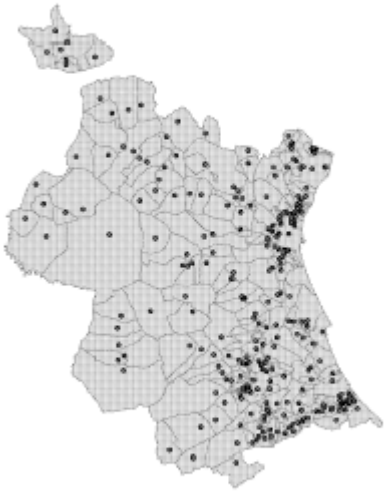
Mapas de mortalidad

- Mortalidad por cáncer de próstata en Valencia
- Datos municipales agregados para el periodo 1975-80
- Los colores indican intensidad del atributo en estudio



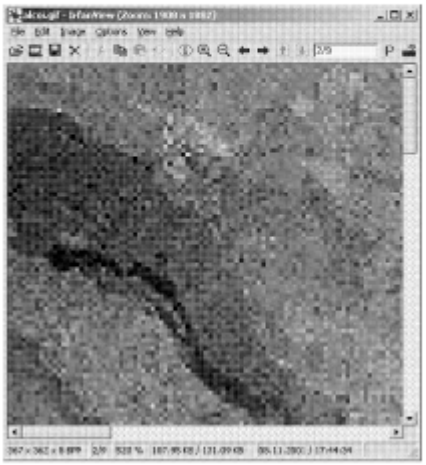
Mapas de mortalidad (2)

- La ubicación exacta suele ser el núcleo de población principal, aunque tambien se usa el baricentro de la poligonal, etc.
- Utilizada para métodos basados en distancias



Imágenes digitales

- Los píxeles constituyen la parrilla de localizadiones fijas
- La imagen verdadera está enmascarada por perturbaciones aleatorias



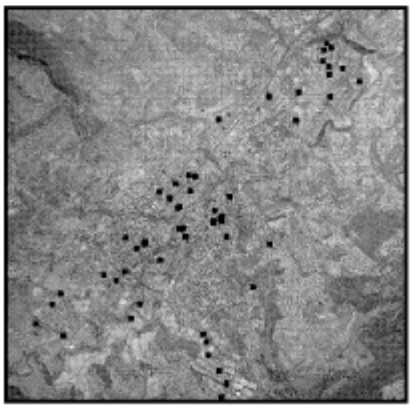
Tipos de datos

DATOS SOBRE

- Localización continua (pluviometría, polución ambiental, ...)
- Parrilla fija (estadísticas municipales, imágenes digitales,...)
- Localizaciones aleatorias (ubicación cráteres, nidos de aves, ...)

Casos de legionela en Alcoy

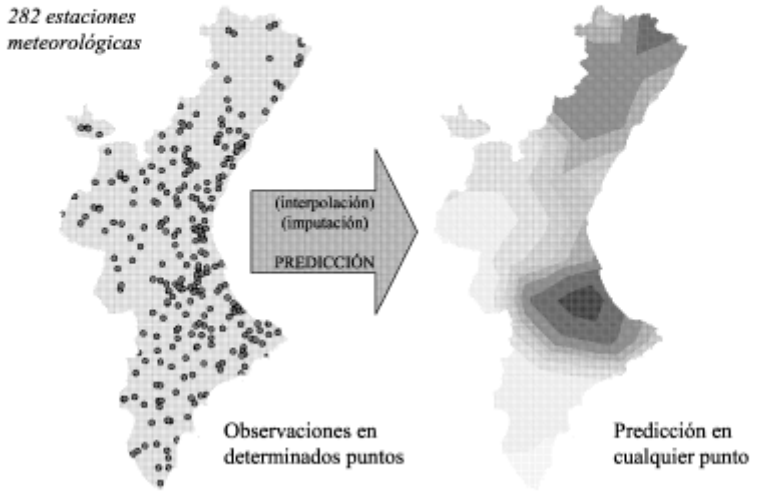
- Residencia como localización aleatoria
- Torres de refrigeración como segundo proceso de ubicaciones aleatorias



CAMPOS ALEATORIOS SOBRE CONTINUO DE LOCALIZACIONES

"KRIGING"

Problema a resolver



Procedimientos

- DETERMINISTAS
  - –Interpolación: la superficie estimada *toca* los valores originales en los puntos de muestra
  - –Suavizado: la regularidad de la superficie no conserva estos valores originales
- ESTADÍSTICOS
  - –KRIGING: desarrollado en el contexto de ingenieria de minas a partir de los 60 como un híbrido de ingeniería, matemáticas y estadística. Matheron le da ese nombre en honor al ingeniero sudafricano D.G. Krige

Formulación del Kriging

- Notación $y_s,\ s \in S \subset \mathbb{R}^2 \implies \hat{y}_r,\ r \in \mathbb{R}^2 \setminus S$
- Modelo $$y_r = \sum_j \beta_j f_j(r) + \varepsilon(r)$$

$$\mathrm{E}[\boldsymbol{\varepsilon}] = 0$$

$$\mathrm{V}[\boldsymbol{\varepsilon}] = \Sigma$$

- Se plantea como problema de *Predicción Lineal Insesgada Óptima*
- Corresponde a un modelo lineal auto-regresivo

Tendencia y variabilidad

- Tendencia $\sum \beta_j f_j(r)$
  - Capta la Variabilidad a gran escala:
    - Kriging simple: tendencia conocida
    - Kriging ordinario: constante desconocida
    - Kriging universal: tendencia polinómica
- Variabilidad a pequeña escala $\varepsilon(r)$
  - Proceso estocástico estacionario intrínseco y posiblemente istrópico

$$V(y_r - y_s) = 2\gamma(|r - s|)$$

Variograma

- Propiedades $\gamma(d) = \frac{1}{2}V[y_0 - y_d]$
  - Creciente con la distancia
  - Decreciendo a cero con *h* tendiendo a cero
  - Se acepta la inclusión de una perturbación aleatoria local o efecto “pepita”
- Modelos válidos: corresponden a funciones *definidas negativas condicionales*

$$\{s_i\}_{i=1}^n \subset \mathbb{R}^2,\ \boldsymbol{a} \in \mathbb{R}^n \quad \Longrightarrow \quad \sum_i \sum_j a_i a_j \gamma(s_i - s_j) \leq 0$$

Partes del variograma

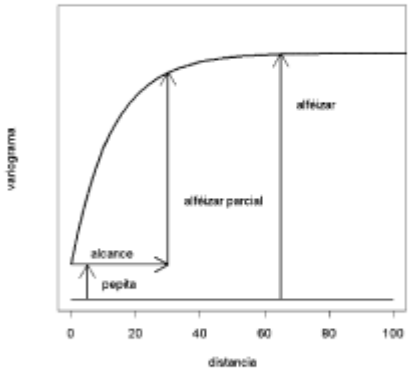
Pepita: varianza puntual local

Alcance: distancia hasta que la influencia se hace muy pequeña

Alféizar: máximo valor del variograma (corresponde a la varianza local)

Alféizar parcial: diferencia entre alféizar y pepita

En inglés: nugget, range, sill, partial sill.



Algunos variogramas isotrópicos

- Exponencial $\gamma(d) = \begin{cases} 0 & d = 0 \\ a + b[1 - \exp(-d/c)] & d \neq 0 \end{cases}$

  $a \geq 0,\ b \geq 0,\ c \geq 0$

- Esférico $\gamma(d) = \begin{cases} 0 & d = 0 \\ a + b\left[\frac{3}{2}(d/c) - \frac{1}{2}(d/c)^3\right] & 0 < d \leq c \\ a + b & d > c \end{cases}$

  $a \geq 0,\ b \geq 0,\ c \geq 0$

- Racional cuadrático $\gamma(d) = \begin{cases} 0 & d = 0 \\ a + \dfrac{bd^2}{1 + d^2/c} & d \neq 0 \end{cases}$

  $a \geq 0,\ b \geq 0,\ c \geq 0$

Estimación del variograma

- Variograma empírico: tras filtrar la tendencia, se calculan las diferencias entre residuos de todos los pares posibles de puntos muestrales
- Ajuste de variograma teórico: se escoge un modelo válido y se ajusta al variograma empírico
- Valoración por tests de Monte Carlo y comparación de modelos alternativos mediante validación cruzada del resultado final

Estimación de la tendencia

- Mínimizar error cuadrático medio

$$\text{mín } E[(y_r - \hat{y}_r)^2]$$

Predictor lineal → $\hat{y}_r = \sum_{s \in S} \lambda_s y_s$

Exigencia de insesgadez → $\boldsymbol{\lambda}' \boldsymbol{f}(s) = \boldsymbol{f}(r)$

- Predictor kriging= predictor m.c.g. + corrección por residuales correlados

$$\hat{\boldsymbol{\beta}} = [F_S' \Sigma_{SS}^{-1} F_S]^{-1} \Sigma_{SS}^{-1} F_S' \boldsymbol{y}_S$$

$$\hat{y}_r = \boldsymbol{f}(r)' \hat{\boldsymbol{\beta}} + \Sigma_{rS} \Sigma_{SS}^{-1} \boldsymbol{e}_S$$

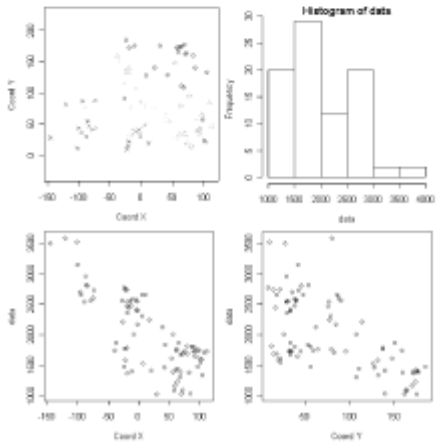
Acuífero de Wolfcamp

- Nevada, Texas: en el subsuelo de una región destinada a enterramiento de residuos radiactivos, el acuífero de agua salada amenaza con la corrosión de la envoltura de los desechos.
- Se midió la presión del acuífero (en pies de altura sobre el nivel del mar) mediante sondeo de pozos

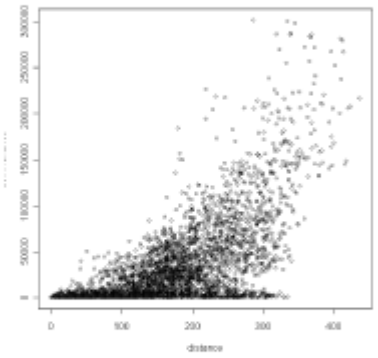
Análisis exploratorio

1. Quintiles de la presión por colores
2. Histograma de valores de la presión
3. Valores de la presión según longitud geográfica
4. Valores de la presión según latitud geogrráfica

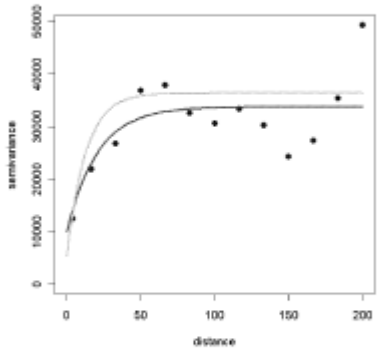
- Gradiente negativo tanto hacia norte como hacia el este



Variograma



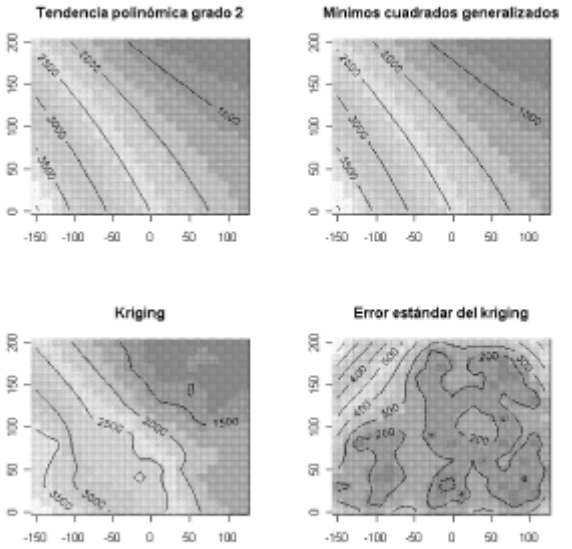
Contribuciones de los pares de puntos



Variograma ajustado

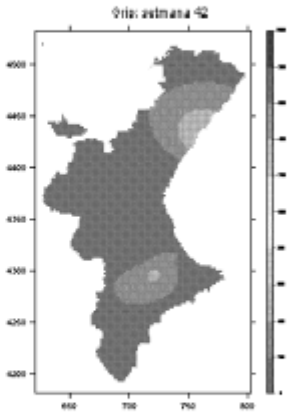
Ajuste m.c.

Predic-ción Kriging



Gripe en la C. Valenciana

- Campaña 1998-99
- Predicción a partir de la red de vigilancia
- Kriging independiente a cada semana

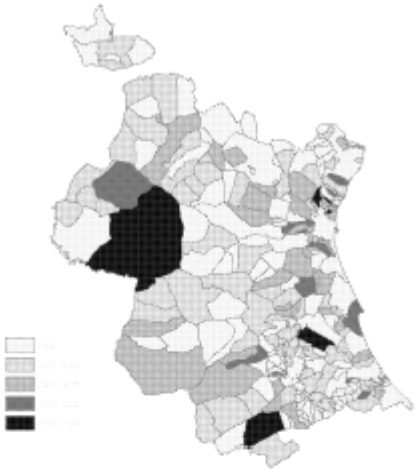


DATOS EN RETÍCULO FIJO DE LOCALIZACIONES

Problema que se plantea

Dada la distribución geográfica de una magnitud aleatoria de interés:

1. Buscar explicación de su variabilidad geográfica (ej. confección de atlas de enfermedades)
2. Verificar influencia de posibles factores de riesgo (ej. contaminación por nitratos del agua de suministro público)

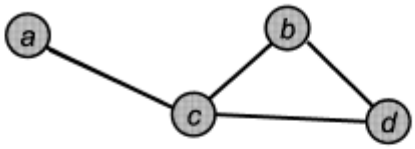


Modelos espaciales

- Autoregresivos: los valores del atributo de interés en localidades vecinas se influencian mutuamente (ejemplo: enfermedades infecciosas en plantas o animales)
- Jerárquicos: la correlación entre valores del atributo en localidades próximas se debe a factores ocultos cuyo alcance supera la dimensión de las unidades geográficas consideradas (ejemplo: polución del aire y enfermedades respiratorias)

Modelos Markovianos: vecindad

- Relación de vecindad



- Interpretación Markoviana del grafo (grafos de independencia)

$$p(y_a \mid y_{-a}) = p(y_a \mid y_c)$$
$$p(y_b \mid y_{-b}) = p(y_b \mid y_c, y_d)$$
$$p(y_c \mid y_{-c}) = p(y_c \mid y_a, y_b, y_d)$$
$$p(y_d \mid y_{-d}) = p(y_d \mid y_b, y_c)$$

Distribuciones de Gibbs

- Especificaciones locales: bajo condiciones muy generales (Teorema de Hammersley-Clifford)

$$\{p(y_i \mid y_{-i})\}_{i \in S} \implies p(\boldsymbol{y})$$

- Función negpotencial: toda distribución de Gibbs se puede expesar

$$p(\boldsymbol{y}) \propto \exp\{U(\boldsymbol{y})\}$$
$$U(\boldsymbol{y}) = \sum_{\delta \in \mathcal{C}} U(\boldsymbol{y}_\delta)$$

interacciones

cliques

Automodelos de Besag

Limitan las especificaciones locales a ditribuciones de la familia exponencial y a interacciones de orden dos

- Autobinomial $y_i \mid \boldsymbol{y}_{-i} \sim \text{Binomial}(n_i, \pi_i)$

$$\log \frac{\pi_i}{1-\pi_i} = \alpha_i + \sum_{j \in \delta(i)} \theta_{ij} y_j$$

- Auto-Poisson $y_i \mid \boldsymbol{y}_{-i} \sim \text{Poisson}(\lambda_i)$

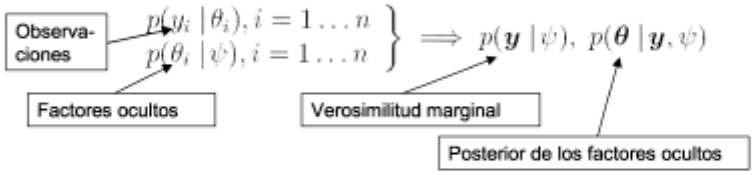
$$\log \lambda_i = \alpha_i + \sum_{j \in \delta(i)} \theta_{ij} y_j$$

Ajuste de los modelos

- Pseudoverosimilitud: considera las especificaciones locales como si fuera de observaciones independientes
- Métodos de Monte Carlo mediante cadenas de Markov (MCMC)
  - Muestreo de Gibbs: se visitan sucesivamente las variables simulando de la especificacion local
  - Se estima por Monte Carlo la verosimilitud, sus derivadas, etc. y puede aplicarse algoritmos de optimización

Modelos jerárquicos

- Estructura de capas: observaciones, factores ocultos, parámetros del modelo



- Transferencia de información: la información aportada por un dato se transfiere a factores asociados a otros datos a través del aprendizaje sobre los parámetros

Análisis Bayesiano y MCMC

- Ciclo de aprendizaje:
  posterior = inicial · verosimilitud / predictiva

$$\left.\begin{array}{l} p(y_i \mid \theta_i), i = 1 \ldots n \\ p(\theta_i \mid \psi), i = 1 \ldots n \\ p(\psi) \end{array}\right\} \Longrightarrow \left\{\begin{array}{l} p(\boldsymbol{\theta} \mid \boldsymbol{y}, \psi) \\ p(\psi \mid \boldsymbol{y}) \end{array}\right.$$

- Muestreo de la posterior mediante MCMC

$$p(x_i \mid \boldsymbol{x}_{-i}) \propto p(\boldsymbol{x})$$
$$p(\psi, \boldsymbol{\theta} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \psi) p(\psi)$$

Modelos autoregresivos versus jerárquicos

- Interacción o factores ocultos? Ambos mecanismos producen observaciones correladas espacialmente. El modelo jerárquico conduce a verosimilitudes más suaves
- Realismo: ambas posibilidades tienen significado en el contexto de aplicación. Todavía está por desarrollar modelos que integren ambas

Cáncer de próstata en Valencia (1975-80)

CÓDIGO	NOMBRE	Población	%>45	Nitratos	C.prost.
4609001	Ademuz	1545	0.59	11	1
4625002	Ador	1256	0.49	16	5
4624003	Adzaneta de Albaida	1364	0.43	18	0
4624004	Agullent	2016	0.36	8	0
4614005	Alaquas	23728	0.32	78	5
4624006	Albaida	5573	0.39	8	3
4616007	Albal	8139	0.36	17	4
4621008	Albalat de la Ribera	3594	0.42	76	2
4613009	Albalat dels Sorells	567	0.41	60	8
...	...	...	...	...	...

C. Próstata: modelo auto-Poisson

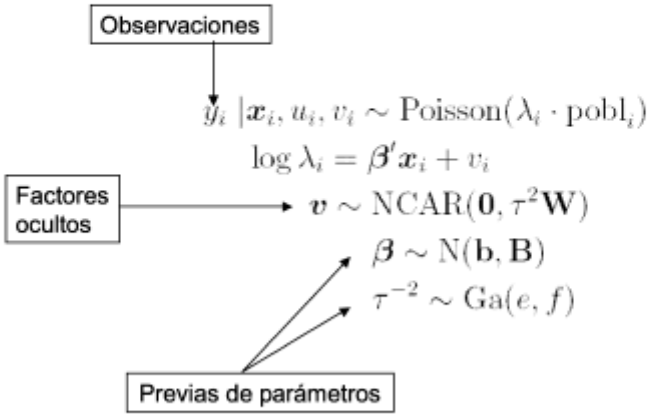
$$\text{casos}_i \sim \text{Poisson}(\lambda_i \times \text{pobl}_i)$$

$$\log(\lambda_i) = \beta_0 + \beta_1 \times \text{envej}_i + \beta_2 \times \text{nitr}_i + \beta_3 \times \sum_{j \sim i} \text{casos}_j$$

Análisis de desvianza	Envejecimiento	Envejecimiento +Autoreg.	Envejecimiento +Autoreg. +Nitratos
Desvianza	361.8	347.1	343.9
+Auto-regresivo	14.7**		
+Nitratos	7.9**	3.2	

Método	Estimación	99%I.C.
Regr. de Poisson	$2.09 \cdot 10^{-3}$	$[0.25, 3.90] \cdot 10^{-3}$
Auto-P (pseudov.)	$1.41 \cdot 10^{-3}$	$[-0.54, 3.36] \cdot 10^{-3}$
Auto-P (MCMC)	$1.50 \cdot 10^{-3}$	$[-0.42, 3.42] \cdot 10^{-3}$

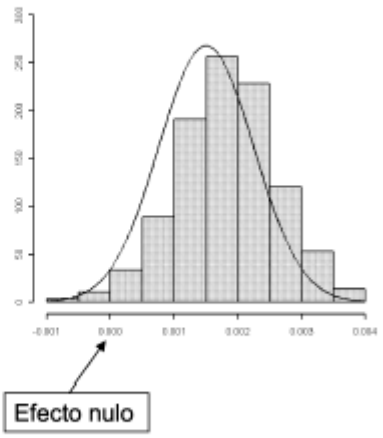
C. Próstata: modelo jerárquico



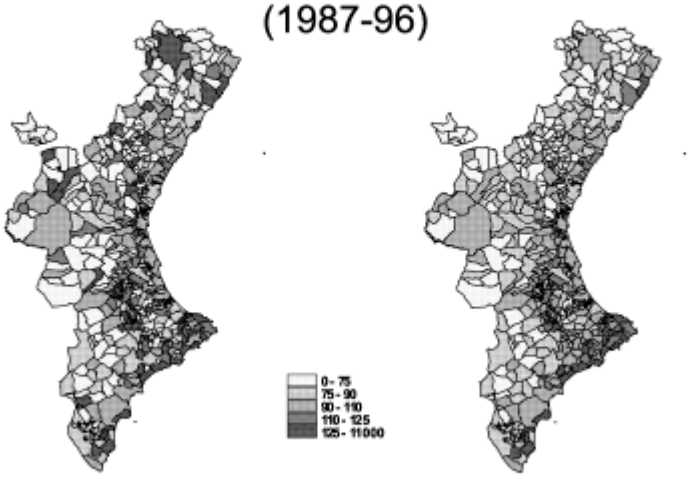
Modelo auto-Poisson vs. Jerárquico

Coeficiente de regresión del nitrato

- Linea contínua: verosimilitud estimada mediante muestreo de Gibbs del modelo auto-Poisson
- Histograma: valores obtenidos por muestreo Gibbs de la posterior del modelo jerárquico.
- Apuntan a la misma conclusión: la evidencia del efecto del nitrato no acaba de ser concluyente



Mortalidad por isquémicas en hombres (1987-96)



Tasas brutas

Tasas suavizadas

LOCALIZACIONES ALEATORIAS: PATRONES PUNTUALES

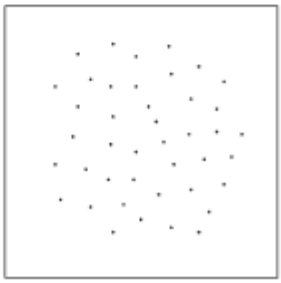
Tipos de patrones

- Completamente aleatorio
- Agregado
- Regular

pinos

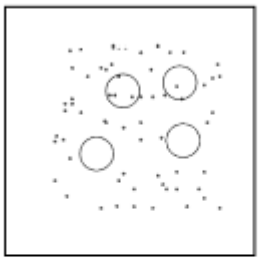
semillas

células



Técnicas básicas

- Recuentos en recintos
  - Aleatorios
  - Prefijados



- Distibuciones de distancias
  - Entre eventos
  - Punto-evento

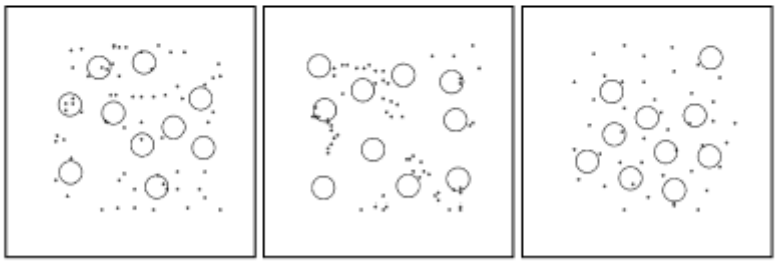
Recuentos en recintos

Índices de Agrupamiento

- **Varianza relativa** $\frac{s^2}{\bar{x}}$
- **David-Moore** $\frac{s^2}{\bar{x}} - 1$
- **Desigualdad de Lloyd** $1 + \frac{s^2}{\bar{x}^2} - \frac{1}{\bar{x}}$
- **Morisita** $\frac{\sum_i x_i(1-x_i)}{n\bar{x}(n\bar{x}-1)}$

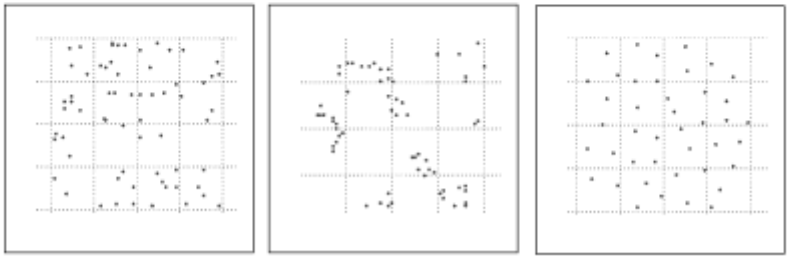
Recintos aleatorios

Ind. agrupamiento	Pinos	Semillas	Células
Varianza relativa	0.9798	2.8120	0.2222
David-Moore	-0.0202	1.8120	-0.7778
Lloyd	0.9844	2.2943	0.1358
Morisita	0.0492	0.1138	0.0065

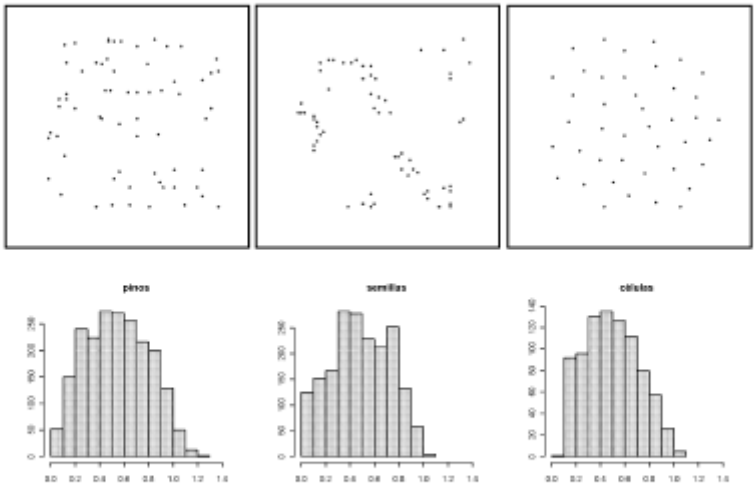


Recintos prefijados

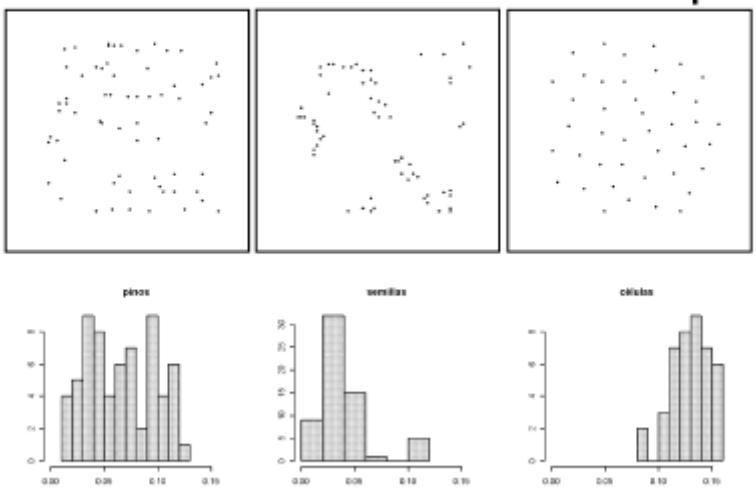
Bondad de ajuste	Pinos	Semillas	Células
Estadístico χ^2	15.0	37.1	2.95
Grados de lib.	15	15	15
P-valor	0.4514	0.0012	0.9996



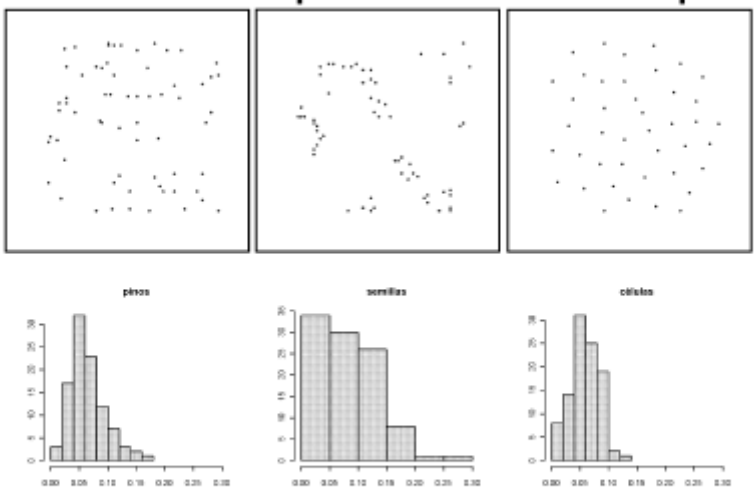
Distancias entre eventos



Distancias evento-evento m.p.



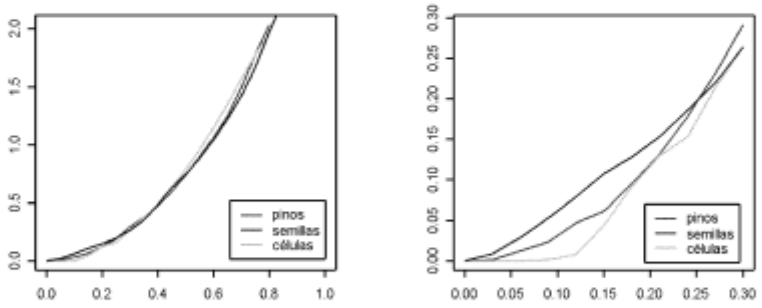
Distancias punto-evento m.p.



Medida momento de orden 2 función $K(r)$

- $K(r)$: número esperado de eventos a distancia no mayor que r de un evento dado, normalizada por la intensidad del proceso.

K(r) empírica

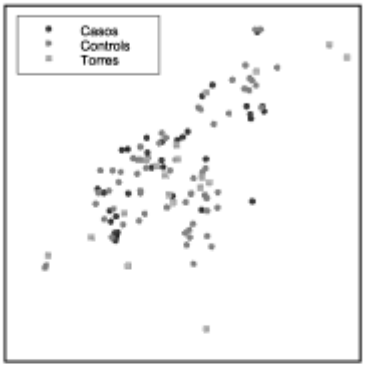


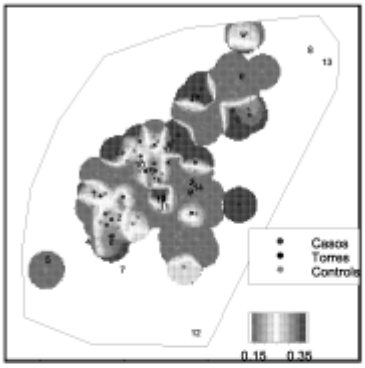
Ajuste y validación de modelos

- Es más fácil simular los modelos de procesos puntuales que obtener teóricamente sus características. Los hay de inhibición y de agregación, “autoregresivos” y “jerárquicos”.
- Ajuste y validación: mediante simulación. Estimación de parámetros y contraste de hipótesis por Monte Carlo.

Brote de legionela en Alcoy

Medida del riesgo:





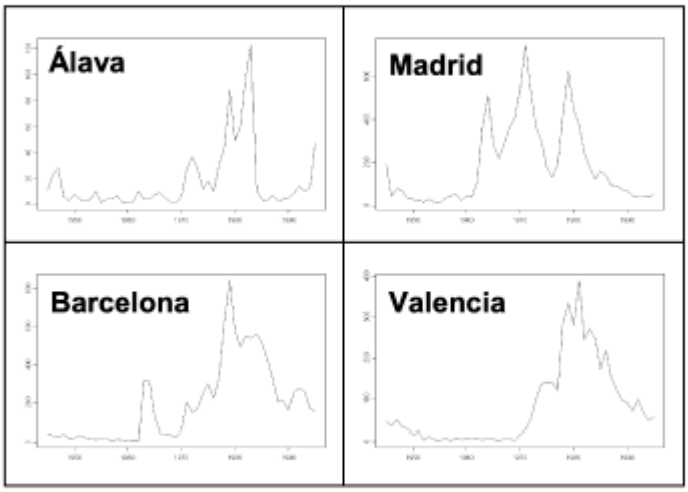
Y PARA TERMINAR...

Temas candentes

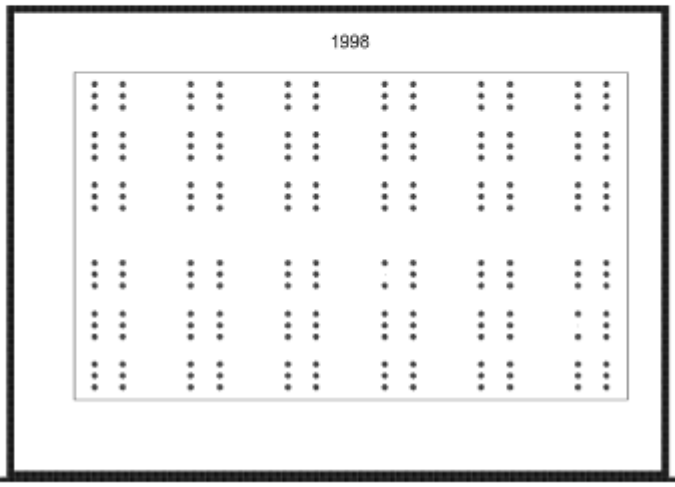
- Análisis espaciotemporal
  - ¿Serie temporal multivariante o concatenación de modelos espaciales?
  - lo más prometedor: modelos de espacios de estado o jerárquicos:
- Sistemas de Información Geográfica (GIS)
  - Enorme interés práctico por su potencial integrador
  - Necesitan incorporar de procedimientos estadísticos

Meningitis 1940

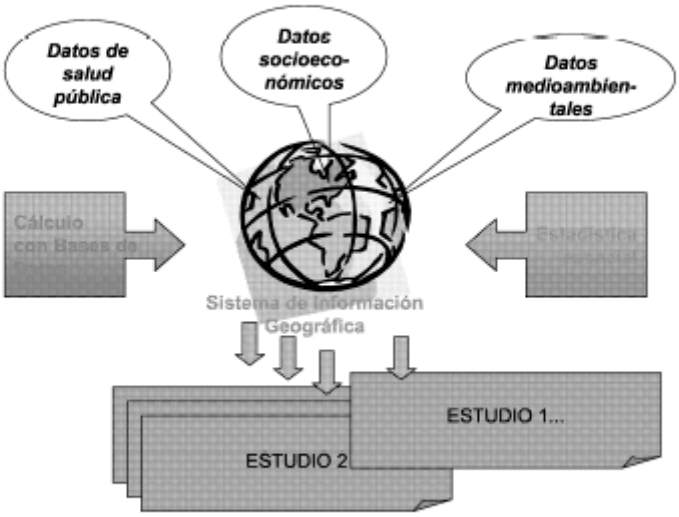
Serie temporal multivariante



Tristeza del naranjo en campo de pruebas del IVIA

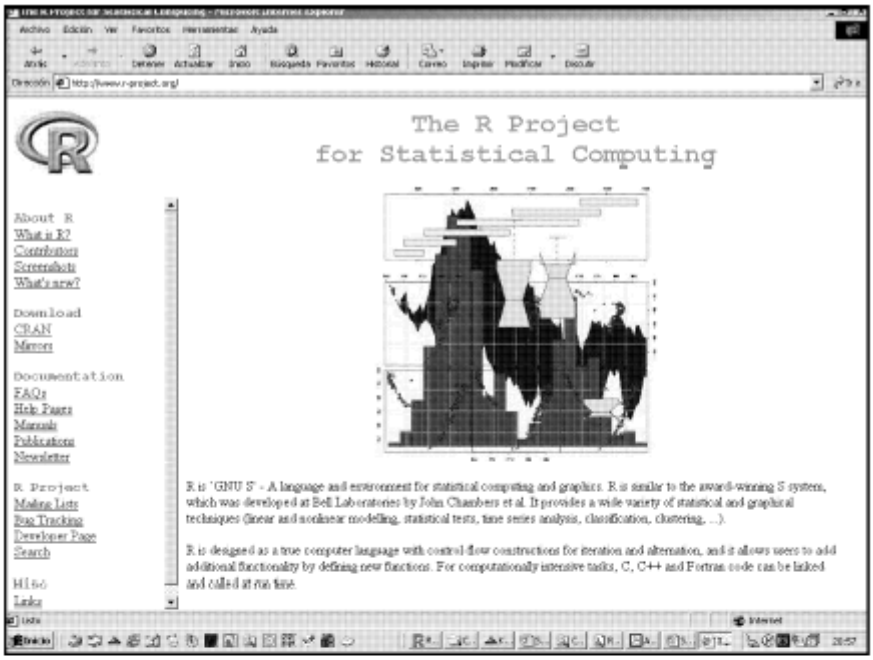


GIS sanitario medioambiental



Software útil

- Entorno R: lenguaje de programación orientado a objetos estadísticos. Excelente calidad y libre distribución http://www.r-project.org
- WinBUGS: muestreo de distribuciones complejas mediante MCMC (ideal para análisis Bayesiano de modelos jerárquicos. Libre distribución http://www.mrc-bsu.cam.ac.uk/bugs/


The R Project
for Statistical Computing
About R
What is R?
Contributors
Screenshots
What's new?
Download
CRAN
Mirrors
Documentation
FAQs
Help Pages
Manuals
Publications
Newsletter
R Project
Mailing Lists
Bug Tracking
Developer Page
Search
Misc
Links
R is 'GNU S' - A language and environment for statistical computing and graphics. R is similar to the award-winning S system, which was developed at Bell Laboratories by John Chambers et al. It provides a wide variety of statistical and graphical techniques (linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, ...).
R is designed as a true computer language with control-flow constructions for iteration and alternation, and it allows users to add additional functionality by defining new functions. For computationally intensive tasks, C, C++ and Fortran code can be linked and called at run time.


The BUGS Project
welcome
Welcome Page
Latest News
Overview & Demo
Discussion list
WinBUGS
GeoBUGS
Classic BUGS
CODA
FAQ Pages
Documentation
About BUGS
Bayesian inference Using Gibbs Sampling is a piece of computer software for the Bayesian analysis of complex statistical models using Markov chain Monte Carlo (MCMC) methods. It grew from a statistical research project at the MRC Biostatistics Unit, but now is developed jointly with the Imperial College School of Medicine at St Mary's, London.
The Classic BUGS program uses text-based model description and a command-line interface, and versions are available for major computer platforms. A Windows version, WinBUGS, has an option of a graphical user interface and has on-line monitoring and convergence diagnostics. CODA is a suite of S-plus/R functions for convergence diagnostics.
The programs are reasonably easy to use and come with a range of examples. Considerable caution is, however, needed in their use, since the software is not perfect and MCMC is inherently less robust than analytic statistical methods. There is no in-built protection against misuse.
About this Site
The new BUGS web site was launched in November 1999 by Alastair Stevens. This site was developed entirely by hand without the use of authoring tools. It is designed to be fully HTML 4.0 compliant, and uses the latest in style sheet techniques. By using style sheets, we've achieved a range of visual effects with the minimum number of images, which makes for very fast-loading pages.

Referencias de las aplicaciones precedentes

Abellán J.J., Vanaclocha H., Zurriaga O., Melchor I.,Calabuig J., Ferrándiz J., López A., y Sanmartín P. (2000). Atlas of cardiovascular mortality in the comunidad Valenciana (1987-1996). {h}ttp://dgsp.san.gva.es/sscc/epidemiologia/estudios/estudios.htm.

Abellán JJ, Martínez-Beneito MA, Zurriaga O, Jorques G, Ferrándiz J, López-Quílez A (2002). Procesos puntuales como herramienta para el análisis de posibles fuentes de contaminación. Pendiente de publicación en *Gaceta Sanitaria.*

Abellán J. J., Óscar Zurriaga, Martínez-Beneito M.A., Peñalver J. y Molins T. Incoporación de la metodología geoestadística para la vigilancia de la gripe en una red centinela. (Submitted)

Ferrándiz J., López A., Llopis A., Morales M.M., Tejerizo M.L (1995) Spatial Interaction between Neighbouring Counties: Cancer Mortality Data in Valencia (Spain). *Biometrics* **51**: 665–678

Ferrándiz J., López A., Sanmartín P. (1999). Spatial regression models in epidemiological studies. *Disease Mapping and Risk Assessment for Public Health Decision Making* (Lawson *et al.* Eds.), pp 203-215. Wiley

Ferrándiz J., Martínez F., Sanmartín P. (2001) Concatenación temporal de modelos espaciales y su aplicación al estudio de la meningitis en España. *Qüestiió* **25**: 47–68.

Ferrándiz J., Abellán J. J., López A., Sanmartín P., Vanaclocha H., Zurriaga O., Martínez-Beneito M. A., Melchor I., Calabuig J. (2002). Geographical distribution of the cardiovascular mortality in Comunidad Valenciana (Spain). *GIS for Emergency preparedness and health risk reduction* (D. Briggs *et al.* eds.). Pendiente de publicación en la editorial Kluwer.

Ferrándiz J., *et al.* (2002). Statistical relationship between hardness of drinking water and cerebrovascular mortality in Valencia: a comparison of spatiotemporal models. Pendiente de publicación en *Environmetrics.*