Instituto Tecnológico Autónomo de México

Construcción y Estimación de Campos Aleatorios Markovianos

Rogelio Ramos Quiroga
Graciela Ma. González Farías
CIMAT

México, D.F., 18 de Mayo, 2007

Temas

- **Aplicaciones de Estadística Espacial.**
- **Construcción de Modelos.**
- **Un Modelo Espacial para Respuestas Ordinales.**
- **Estimación vía Pseudoverosimilitud.**
- **Estimación MV vía Simulación.**
- **Una Comparación de Métodos de Estimación.**
- **Conclusiones.**

Aplicaciones de Estadística Espacial

Aplicaciones en Epidemiología

Mapeo del riesgo de una enfermedad. Tasas de incidencia de cáncer de tiroides, en Francia, en el período 1971–1978.

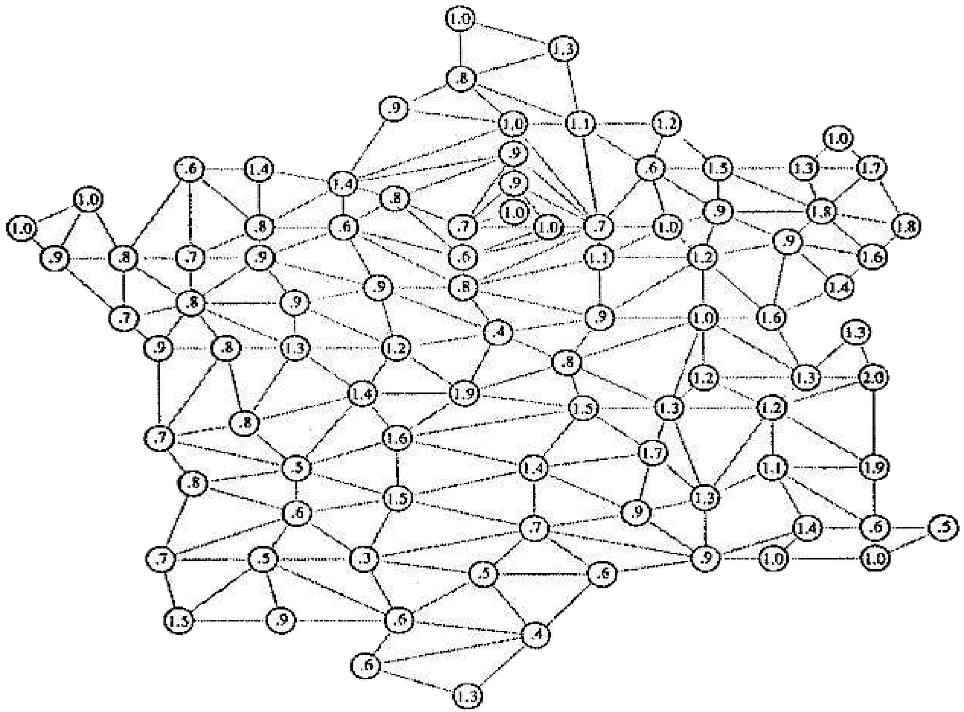
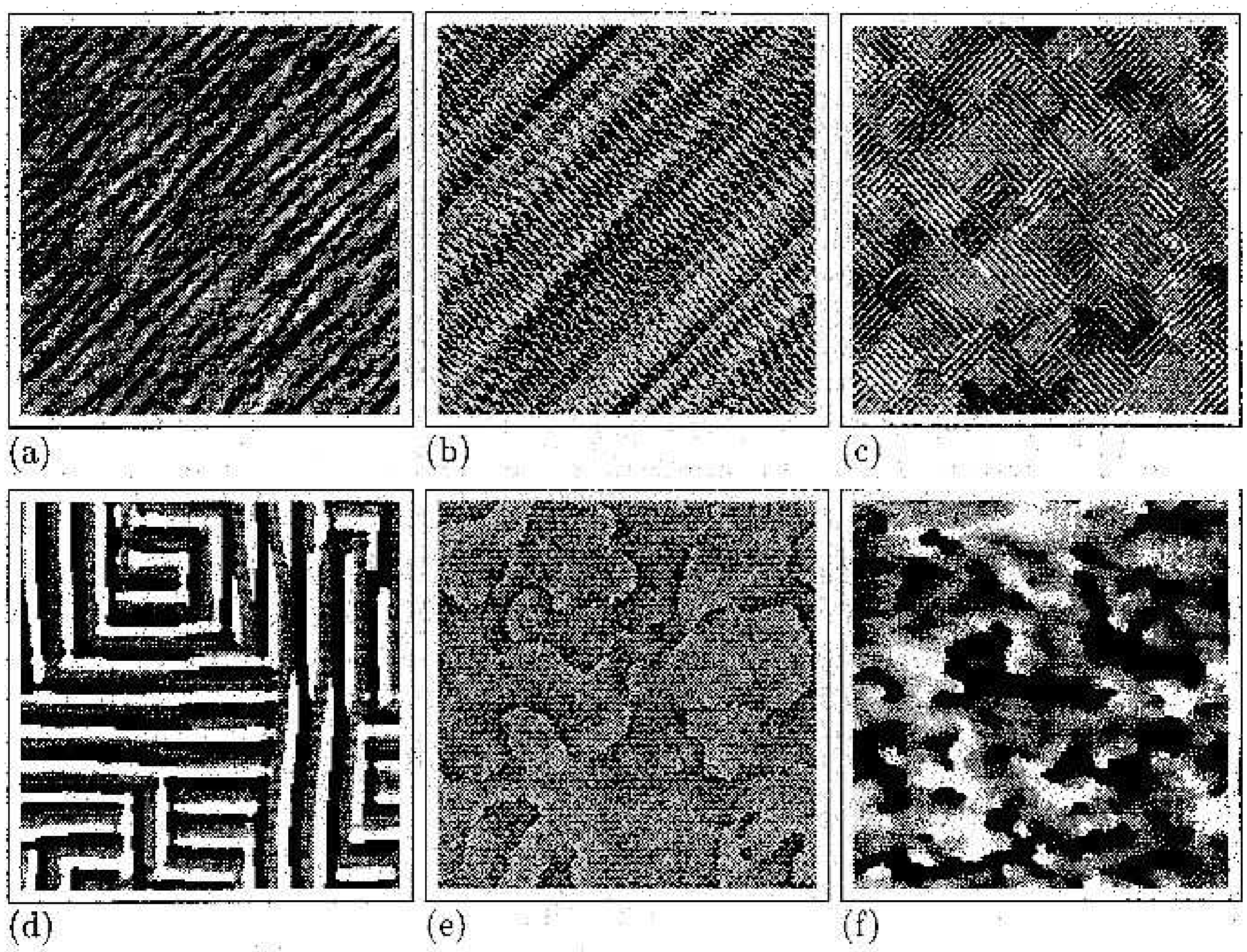


Fig. 4. Observed mortality from thyroid cancer, relative to the overall mean rate.

(Besag, York & Mollié (1991))

Aplicaciones en Procesamiento de Imágenes: Texturas



(Winkler (2003), Cross & Jain (1983), Geman & Geman (1984))

Aplicaciones en Agronomía

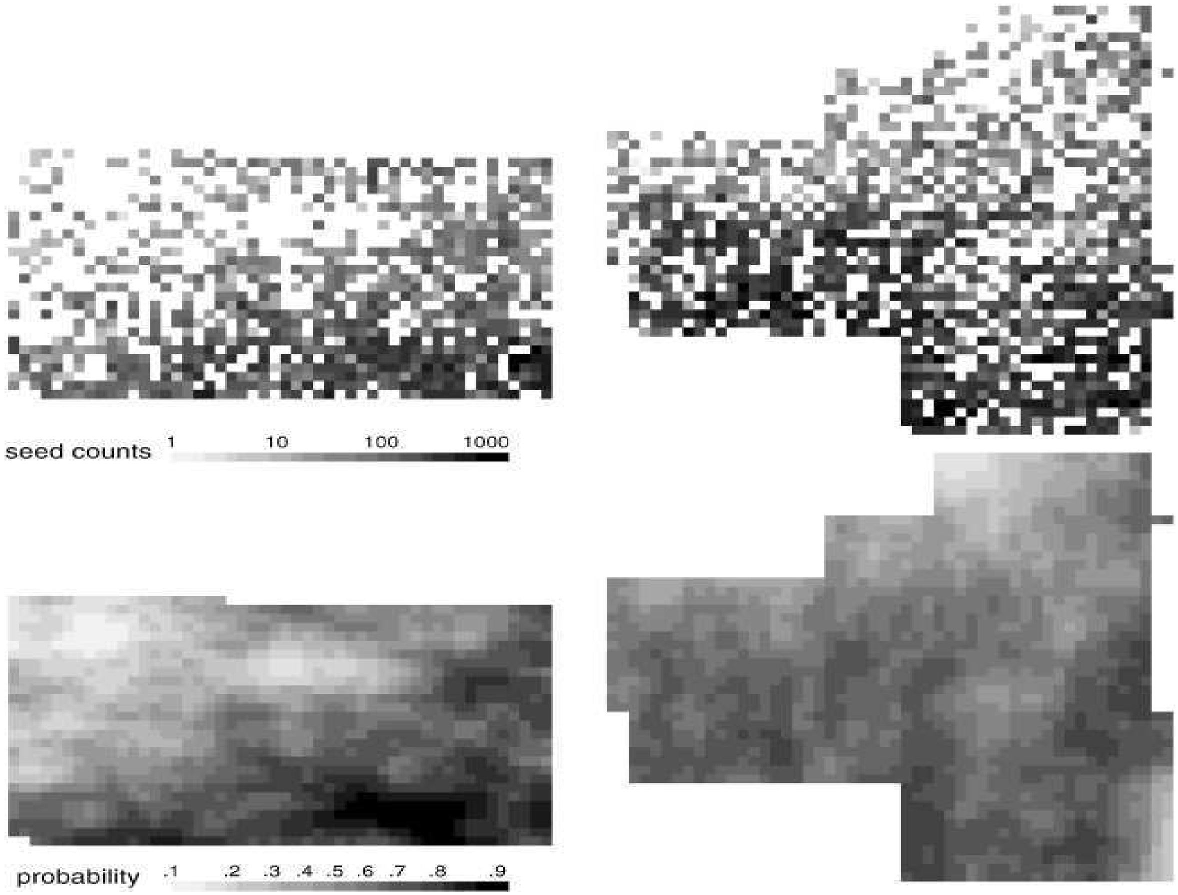


Fig. 3. Raw data (top) and posterior fertility map (bottom) in the morning-glory seed trial: for the raw data, each pixel represents the number of seeds collected from a single plant — the darker the pixel, the higher the count — white pixels identify plots in which no seed was produced; fertilities are shown for all locations where seeds were planted

(Besag & Higdon (1999))

Construcción de Modelos

Modelos Espaciales

Sea $y = (y_1, \cdots, y_n)$, donde y_i es la observación en la $i-$ésima localidad. Una forma natural de modelación es mediante la consideración de relaciones condicionales:

$$P(y_i \mid y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_n)$$

Dado un conjunto de condicionales, la distribución conjunta, $P(y)$, puede obtenerse mediante un teorema de factorización (Brook, (1964)).

Sea $z = (z_1, \cdots, z_n)$ otro conjunto de posibles valores para y, entonces

$$P(y) = P(z) \prod_{i=1}^{n} \frac{P(y_i \mid y_1, \cdots, y_{i-1}, z_{i+1}, z_n)}{P(z_i \mid y_1, \cdots, y_{i-1}, z_{i+1}, z_n)}$$

Modelos Espaciales

$$P(y) = P(z) \prod_{i=1}^{n} \frac{P(y_i \mid y_1, \cdots, y_{i-1}, z_{i+1}, z_n)}{P(z_i \mid y_1, \cdots, y_{i-1}, z_{i+1}, z_n)}$$

De aquí surge un par de problemas al especificar una conjunta mediante condicionales:

- Consistencia, existencia de la distribución conjunta para todas las localidades.

- Cálculo de la constante de normalización (también llamada Función de Partición)

$$P(z) = \left(\sum_{y_1} \cdots \sum_{y_n} \prod_{i=1}^{n} \frac{P(y_i \mid y_1, \cdots, y_{i-1}, z_{i+1}, z_n)}{P(z_i \mid y_1, \cdots, y_{i-1}, z_{i+1}, z_n)} \right)^{-1}$$

Modelos Espaciales

Para asegurar consistencia, la distribución conjunta debe obedecer el Teorema de Hammersley–Clifford, el cual pide que las condicionales completas sean de naturaleza local y entonces

$$P(y) = \frac{1}{C(\theta)} \exp\left\{ \sum_i y_i G_i(\cdot) + \sum_{i<j} y_i y_j G_{ij}(\cdot) + \sum_{i<j<k} y_i y_j y_k G_{ijk}(\cdot) + \cdots + y_1 \cdots y_n G_{1\cdots n}(\cdot) \right\}$$

donde las funciones G son tales que $G_{i_1,\cdots,i_s}$ depende solamente de $y_{i_1}, \cdots, y_{i_s}$, y pueden ser no nulas sólo si $i_1, \cdots, i_s$ forman un clique.

Se dice que el sitio j es vecino del sitio i si $P(y_i \mid y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_n)$ depende de y_j. Sea N_i el conjunto de vecinos de i, una distribución es un Campo Aleatorio Markoviano si

$$P(y_i \mid y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_n) = P(y_i \mid N_i), \quad i = 1, \cdots, n$$

Un *clique* es un conjunto de sitios en los que todos sus elementos son vecinos entre sí.

Modelos Espaciales

El problema general de construcción de modelos condicionales consistentes, fué ampliamente estudiado en los 70's

- Dobruschin (1968)
- Spitzer (1971)
- Hammersley and Clifford (1971)
- Besag (1972)
- Grimmett (1973)
- Besag (1974)
- Strauss (1975)
- Strauss (1977)

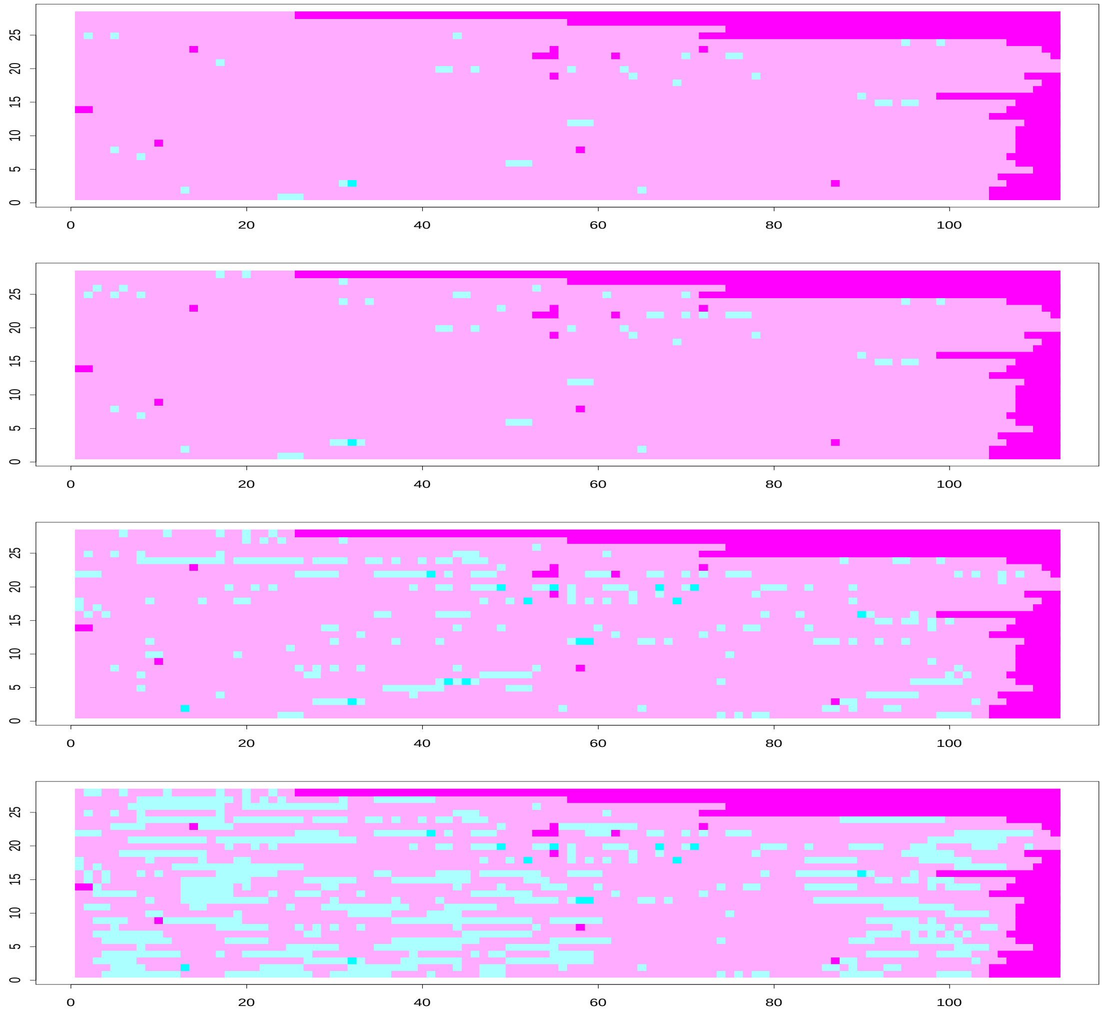
más recientemente, han habido algunas nuevas propuestas de modelos para campos markovianos

- Cressie and Lele (1992)
- Kaiser and Cressie (2000)

Un Modelo Espacial para Respuestas Ordinales

Un Modelo Espacial para Respuestas Ordinales

Mapeo de una enfermedad en plantas de Agave.



Un Modelo Espacial para Respuestas Ordinales

Uno de los modelos más usados con especificaciones locales consistentes, es el modelo espacial de segundo orden (*auto–modelo*) con efectos simétricos para las diferentes clases de vecinos.

$$P(y_i \mid N_i) \propto \exp\left\{ y_i(\alpha_0 + \sum_{k=1}^{d} \beta_k \sum_{t \underset{k}{\sim} i} y_t) \right\}$$

Para respuestas binarias este es el modelo de Ising. Para respuestas continuas, el modelo estándar es el modelo auto–normal con condicionales gaussianas. Consideraremos un modelo de la forma

$$P(y_i \mid N_i) \propto \exp\left\{ y_i(G(y_i) + \sum_{k=1}^{d} \beta_k \sum_{t \underset{k}{\sim} i} y_t) \right\}$$

Donde G es una función cuadrática que permite distribuciones marginales no necesariamente unimodales.

Justificación del Modelo Propuesto

El modelo más general de segundo orden es

$$P(y) = \frac{1}{C(\theta)} \exp \left\{ \sum_i y_i G(y_i) + \sum_{i<j} y_i y_j G(y_i, y_j) \right\}$$

Se puede ver que las funciones G satisfacen

$$y_i G(y_i) = \log \frac{P(y_i \mid \text{resto} = 0)}{P(y_i = 0 \mid \text{resto} = 0)},$$

modelar estos logodds como una función lineal de y_i pudiera ser restrictivo; funciones cuadráticas tendrían comportamiento similar. Un modelo más flexible es

$$G(y_i) = \alpha_1 + \alpha_2 y_i + \alpha_3 y_i^2$$

Tasas de Momios y Parámetros del Modelo

Por otro lado, se puede ver que los términos de segundo orden satisfacen:

$$y_i y_j G(y_i, y_j) = \log \left\{ \frac{P(y_i = r \mid y_j = s)}{P(y_i = 0 \mid y_j = s)} \div \frac{P(y_i = r \mid y_j = 0)}{P(y_i = 0 \mid y_j = 0))} \right\}$$

entonces, los términos $y_i y_j G(y_i, y_j)$ están dados por el log de las tasas de momios correspondientes a una tabla de contingencia $(S+1) \times (S+1)$, (cada localidad con posibles valores $0, 1, \cdots, S$).

Tasas de Momios y Parámetros del Modelo

El modelo loglineal más simple para una tabla $(S+1)\times(S+1)$ que toma en cuenta la ordinalidad de las respuestas está dado por

$$\log m_{rs} = \mu + \lambda_r^{y_i} + \lambda_s^{y_j} + \beta(r-\bar{r})(s-\bar{s})$$

donde m_{rs} es el valor esperado para la celda (r,s), esto es, $m_{rs} = NP(y_i = r, y_j = s)$, sea $\pi_{rs} = P(y_1 = r, y_2 = s)$. Para este modelo, se puede ver que el log de la razón de momios para las celdas (r,s) y (t,u) es

$$\log \frac{\pi_{rs}\pi_{tu}}{\pi_{ru}\pi_{ts}} = \beta(r-t)(s-u)$$

ahora, como $y_i y_j G(y_i, y_j)$ es simplemente el log de la razón de momios de las celdas (r,s) y $(0,0)$ entonces un modelo razonable para este término sería βrs. Esto es,

$$y_i y_j G(y_i, y_j) = \beta y_i y_j.$$

Distribuciones Condicionales

La distribución condicional en cada sitio, dado sus vecinos, es un miembro de la familia Exponencial

$$P(y_i \mid N_i) = a_0(\theta) t_0(y_i) \exp\left\{\theta^T t(y_i)\right\}$$

donde

$$\begin{aligned}
a_0(\theta) &= \left[\sum_{y_i} \exp\left\{\theta^T t(y_i)\right\}\right] \\
t(y_i) &= \left(y_i, y_i^2, y_i^3, y_i \sum_{j_1} y_{j_1}, y_i \sum_{j_2} y_{j_2}\right)^T \\
t_0(y_i) &= 1
\end{aligned}$$

Inferencia

Máxima Pseudoverosimilitud

Inferencia basada en la verosimilitud para

$$P(y) = \frac{1}{C(\theta)} \exp\left\{ \sum_i y_i G(y_i) + \sum_{i<j} y_i y_j G(y_i, y_j) \right\}$$

implica el cálculo de la *constante normalizadora* $C(\theta)$, lo cual, para un problema pequeño con un látice 10×10 con 3 posibles estados en cada localidad, requeriría la suma de $3^{100} \doteq 5 \times 10^{47}$ términos.

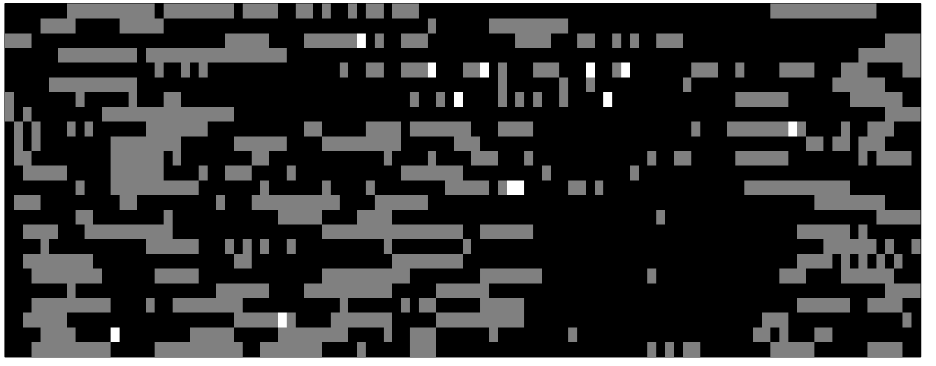
Besag (1974) introdujo los *métodos de estimación basados en la pseudoverosimilitud*, los cuales maximizan la pseudoverosimilitud (evitando con ello calcular $C(\theta)$)

$$PL(\theta) = \sum_{i=1}^{n} \log P(y_i \mid N_i)$$

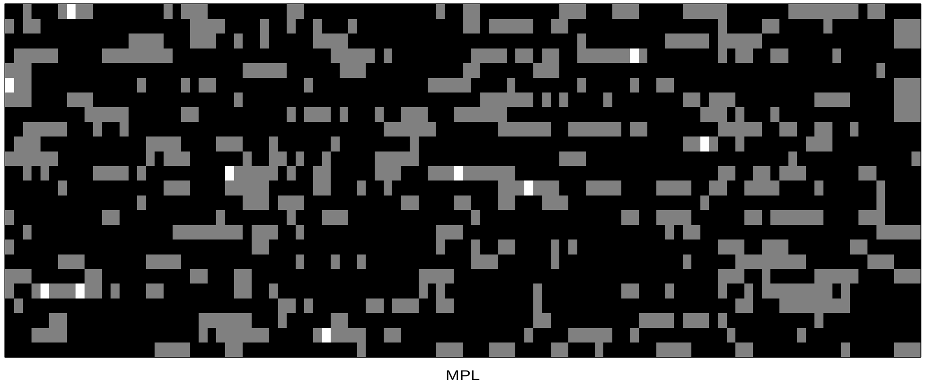
(**Nota: Besag se ha pronunciado activamente por el uso de métodos de estimación más modernos.**)

Resultados sobre consistencia y normalidad asintótica para el *EMPV* pueden encontrarse en Guyon (1995).

Máxima Pseudoverosimilitud



Data



MPL

Estimaciones *MPV*

$$
\begin{aligned}
\widehat{\alpha}_1 &= -5.36 \\
\widehat{\alpha}_2 &= 6.03 \\
\widehat{\alpha}_3 &= -11.7 \\
\widehat{\beta}_{11} &= 9.33 \\
\widehat{\beta}_{12} &= 0.60 \\
\widehat{\beta}_{21} &= -0.39 \\
\widehat{\beta}_{22} &= -0.60
\end{aligned}
$$

Distribuciones marginales

	0	1	2
$Datos$	.702	.293	.005
MPV	.740	.256	.004

Máxima Verosimilitud vía Simulación

Resultados Básicos

Para el modelo

$$f(y \mid \theta) = \exp\{-Q(y,\theta) - \log(C(\theta))\}$$

la logverosimilitud es

$$l(\theta) = -Q(y,\theta) - \log(C(\theta))$$

entonces, una iteración *Newton–Raphson* para el *EMV* es

$$\theta^{(k+1)} = \theta^{(k)} - \left[\nabla^2 l(\theta^{(k)})\right]^{-1} \nabla l(\theta^{(k)})$$

La clave de los métodos que permiten maximizar la verosimilitud se basan en observar que $\nabla l(\theta)$ y $\nabla^2 l(\theta)$ pueden expresarse como esperanzas de ciertas cantidades.

Resultados Básicos

$$l(\theta) = -Q(y,\theta) - \log(C(\theta))$$

Observe que

$$\frac{\partial l(\theta)}{\partial \theta} = -\frac{\partial Q(y,\theta)}{\partial \theta} - \frac{\partial \log(C(\theta))}{\partial \theta}$$

tomando esperanzas en ambos lados y observando que la esperan–
za de la función score es cero:

$$\nabla \log(C(\theta)) = \; -\mathsf{E}\left[\nabla Q(y,\theta)\right]$$

De modo que

$$\nabla l(\theta) = -\nabla_{\theta} Q(y,\theta) + \mathsf{E}\left[\nabla_{\theta} Q(y,\theta)\right]$$

Resultados Básicos

$$l(\theta) = -Q(y,\theta) - \log(C(\theta))$$

Observe, por otro lado, que

$$\frac{\partial^2 l(\theta)}{\partial\theta\partial^T\theta} = -\frac{\partial^2 Q(y,\theta)}{\partial\theta\partial^T\theta} - \frac{\partial^2 \log(C(\theta))}{\partial\theta\partial^T\theta}$$

tomando esperanzas en ambos lados y observando que

$$\mathsf{E}\left(\frac{\partial^2 l(\theta)}{\partial\theta\partial^T\theta}\right) = -\mathsf{E}\left(\frac{\partial l(\theta)}{\partial\theta}\frac{\partial l(\theta)}{\partial\theta^T}\right)$$

obtenemos

$$\nabla^2 \log(C(\theta)) = -\mathsf{E}(\nabla^2 Q(y,\theta)) + \mathsf{E}(\nabla l(\theta)\nabla^T l(\theta))$$

De modo que

$$\nabla^2 l(\theta) = -\nabla^2 Q + \mathsf{E}(\nabla^2 Q) - \mathsf{E}(\nabla Q \nabla^T Q) + \mathsf{E}(\nabla Q)\mathsf{E}(\nabla^T Q)$$

Estimación MV vía Simulación

Para el modelo exponencial, $Q(y, \theta) = -h^T(y)\theta$, donde h es el vector de estadísticos suficientes para θ. Entonces

$$\nabla Q(y, \theta) = -h(y), \quad \nabla^2 Q(y, \theta) = 0$$

Usando las expresiones anteriores, la iteración *Newton–Raphson* se reduce a

$$\theta^{(k+1)} = \theta^{(k)} + (\mathsf{V}[h(y)])^{-1} (h(y) - \mathsf{E}[h(y)])$$

donde

$$\mathsf{V}[h(y)] = \mathsf{E}\left[h(y)h^T(y)\right] - \mathsf{E}\left[h(y)\right] \mathsf{E}^T \left[h(y)\right]$$

Conceptualmente simple, **pero** computacionalmente difícil, debido a que todas las esperanzas involucran la constante normalizadora $C(\theta^{(k)})$.

Estimación MV vía Simulación

Sin embargo, si pudieramos simular $y_1, y_2, \cdots, y_N$ de $f(y \mid \theta^{(k)})$, entonces podríamos aproximar todas las esperanzas involucradas en

$$\theta^{(k+1)} = \theta^{(k)} + (\mathsf{V}[h(y)])^{-1} (h(y) - \mathsf{E}[h(y)])$$

por ejemplo

$$\mathsf{E}\,[h(y)] = \int h(y) f(y \mid \theta^{(k)}) dy \approx \frac{1}{N} \sum^{N} h(y_i)$$

Típicamente, la simulación es efectuada mediante el algoritmo *Metropolis–Hastings* o usando el *Gibbs Sampler*.

Estimación MV vía Simulación

No hay muchas implementaciones directas del *Newton–Raphson Monte Carlo*.

- Huang y Ogata (1999). Efectúan solo una iteración, iniciando en el estimador de máxima pseudoverosimilitud.

- Gu y Zhu (2001). Hacen iteraciones completas hasta convergencia
$$\theta^{(k+1)} = \theta^{(k)} - \gamma_k \left[\nabla^2 l(\theta^{(k)})\right]^{-1} \nabla l(\theta^{(k)})$$
Las cantidades γ_k's son constantes que controlan la longitud del salto (usan dos fases, en la primera con constantes relativamente grandes y al acercarse a convergencia cambian a constantes más pequeñas).

Aproximaciones a la Verosimilitud

Geyer y Thompson (1992), Geyer (1994) propusieron aproximaciones Monte Carlo a la verosimilitud completa

$$l(\theta) = -Q(y,\theta) - \log(C(\theta)) = h^T(y)\theta - \log(C(\theta))$$

Note que

$$f(y|\theta) = \frac{1}{C(\theta)} e^{h(y)^T\theta} \quad \text{y} \quad C(\theta) = \int e^{h(y)^T\theta} dy$$

Si ψ es una parámetro fijo, entonces

$$l(\theta) - l(\psi) = h^T(y)\theta - h^T(y)\psi - \log \frac{C(\theta)}{C(\psi)}$$

$$l(\theta) = c + h^T(y)\theta - \log \frac{C(\theta)}{C(\psi)} = c + h^T(y)\theta - \log \int \frac{e^{h(y)^T\theta}}{C(\psi)} dy$$

Aproximaciones a la Verosimilitud

$$l(\theta) = c + h^T(y)\theta - \mathsf{log} \int \frac{e^{h(y)^T\theta}}{C(\psi)} dy$$

$$l(\theta) = c + h^T(y)\theta - \mathsf{log} \int e^{h^T(y)(\theta-\psi)} f(y \mid \psi) dy$$

Entonces, la simulación de $f(y \mid \psi)$ da la aproximación

$$l_N(\theta) = h^T(y)\theta - \mathsf{log} \left[\frac{1}{N} \sum^N \mathsf{exp} \left[h^T(y_i)(\theta - \psi) \right] \right]$$

El maximizador de $l_N(\theta)$ es el estimador de *Verosimilitud Monte Carlo* de Geyer y Thompson [**Fuertemente promocionado por Geyer (1999)**].

Gradiente Estocástico

La idea más básica para maximizar la verosimilitud es mediante métodos del gradiente

$$\begin{aligned}\theta^{(k+1)} &= \theta^{(k)} + \gamma_k \nabla l(\theta^{(k)}) \\ \theta^{(k+1)} &= \theta^{(k)} + \gamma_k \left(h(y) - \mathsf{E}[h(y)]\right)\end{aligned}$$

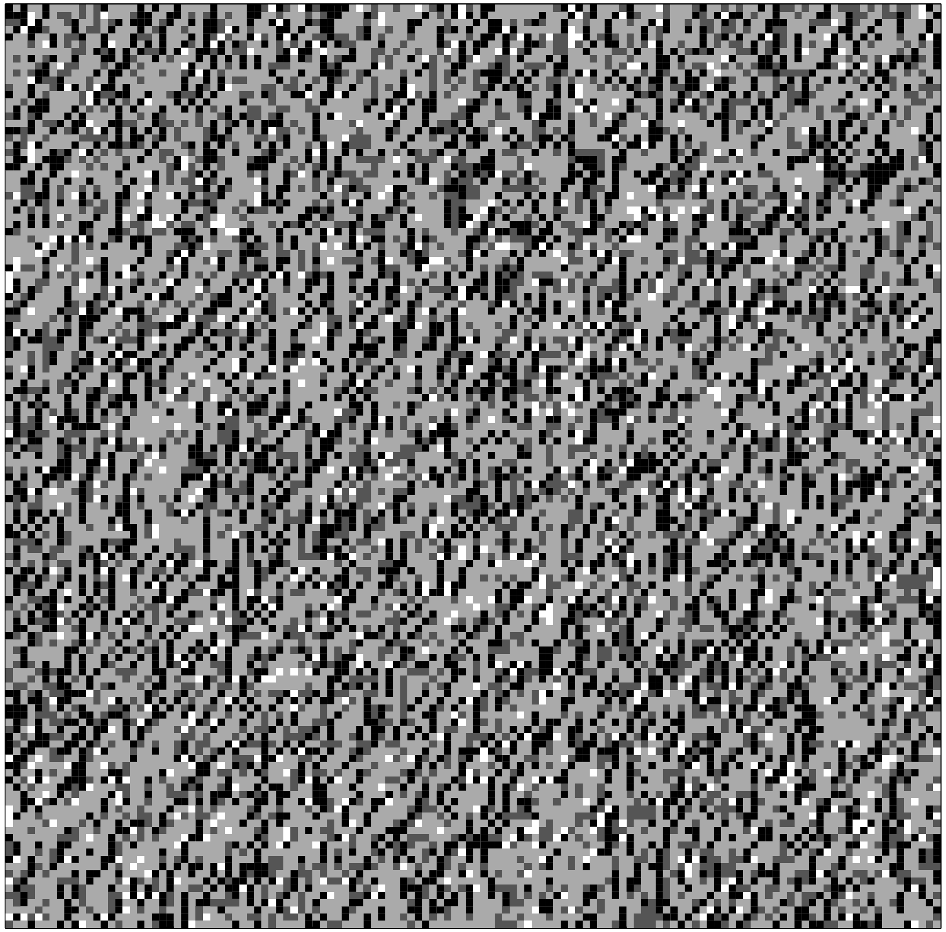
donde, típicamente, $\sum \gamma_k = \infty$ y $\sum \gamma_k^2 < \infty$ ("ni muy lento, ni muy rápido").

- Younes (1991)
- Moyeed y Baddeley (1991)
- Winkler (2001)

(Basados en aproximaciones estocásticas del tipo *Robbins–Monro*).

Una Comparación de Métodos de Estimación

Ejemplo de Prueba



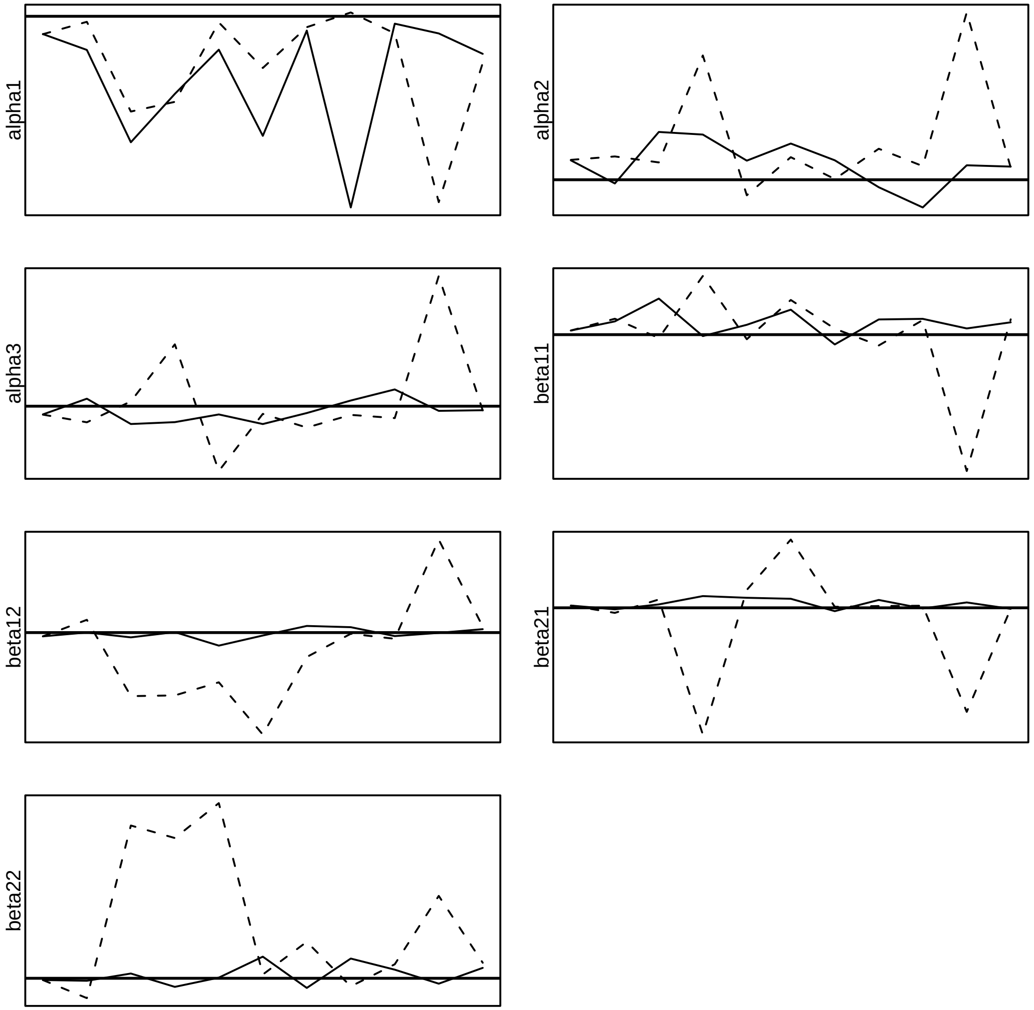
Modelo para texturas con algo de interacciones diagonales.

$$
\begin{aligned}
\alpha_1 &= -9.0 \\
\alpha_2 &= 24.5 \\
\alpha_3 &= -21.0 \\
\beta_{11} &= 4.0 \\
\beta_{12} &= -4.0 \\
\beta_{21} &= 4.0 \\
\beta_{22} &= 0.0
\end{aligned}
$$

Distribución marginal

0	1	2	3
.25	.19	.51	.05

Comparación



Huang–Ogata (línea sólida)

Geyer–Thompson (línea punteada)

Línea de referencia en valores verdaderos.

Simulación de pequeña escala con 10 conjuntos de datos, 2000 muestras sistemáticas vía el Gibbs sampler

Conclusiones

- Presentamos un modelo flexible para datos ordinales espaciales, con posibilidad de representar diferentes patrones de dependencias.

- Estimación de máxima pseudoverosimilitud es un procedimiento computacionalmente estable para modelos espaciales.

- Se dispone de procedimientos para máxima verosimilitud pero su implementación está lejos de ser automática.

- Un estudio limitado de simulación sugiere que el procedimiento *NR* de un paso propuesto por Huang y Ogata podría ser más estable que la aproximación de la verosimilitud de Geyer y Thompson.

Algunas referencias

1. Besag, J. E. (1972). Nearest–neighbour systems and the auto–logistic model for binary data. *Journal of the Royal Statistical Society. Series B*, **34**, 1, 75–83.

2. Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society. Series B*, **36**, 2, 192–236.

3. Besag, J. E., York, J. and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, **43**, 1, 1–59.

4. Besag, J. E. and Higdon, D. (1999). Bayesian analysis of agricultural field experiments. *Journal of the Royal Statistical Society. Series B*, **61**, 4, 691–746.

5. Brook, D. (1964). On the distinction between the conditional probability and the joint probability approaches in the specification of nearest–neighbour systems. *Biometrika*, **51**, 3/4, 481–483.

6. Cressie, N. and Lele, S. (1992). New models for Markov random fields. *Journal of Applied Probability*, **29**, 877–884.

7. Cross, G. R. and Jain, A. K. (1983). Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **5**, 1, 25–39.

8. Dobruschin, P. L. (1968). The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory of Probability and its Applications*, **13**, 2, 197–224.

9. Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

10. Geyer, C. J. (1992). On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society. Series B*, **56**, 1, 261–274.

11. Geyer, C. J. (1999). Likelihood inference for spatial point processes. In *Stochastic Geometry, Likelihood and Computation*, (O. E. Barndorff–Nielsen, W. S. Kendall and M. N. M. van Lieshout, Editors), 79–140. Chapman & Hall.

12. Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *Journal of the Royal Statistical Society. Series B*, **54**, 3, 657–699.

13. Grimmett, G. R. (1973). A theorem about random fields. *Bulletin of the London Mathematical Society*, **5**, 81–84.

14. Gu, M. G. and Zhu, H. T. (2001). Maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation. *Journal of the Royal Statistical Society. Series B*, **63**, 2, 339–355.

15. Guyon, X. (1995). *Random Fields on a Network. Modelling, Statistics, and Applications*. Springer–Verlag.

16. Hammersley, J. M. and Clifford, P. (1971). Markov fields on finite graphs and lattices. (Unpublished).

17. Huang, F. and Ogata, Y. (1999). Improvements of the maximum pseudo–likelihood estimators in various spatial statistical models. *Journal of Computational and Graphical Statistics*, **8**, 3, 510–530.

18. Kaiser, M. S. and Cressie, N. (2000). The construction of multivariate distributions from Markov random fields. *Journal of Multivariate Analysis*, **73**, 199–220.

19. Moyeed, R. A. and Baddeley, A. J. (1991). Stochastic approximation of the MLE for a spatial point pattern. *Scandinavian Journal of Statistics*, **18**, 39–50.

20. Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, **22**, 3, 400–407.

21. Spitzer, F. (1971). Markov random fields and Gibbs ensembles. *American Mathematical Monthly*, **78**, 142–154.

22. Strauss, D. J. (1975). A model for clustering. *Biometrika*, **62**, 2, 467–475.

23. Strauss, D. J. (1977). Clustering on coloured lattices. *Journal of Applied Probability*, **14**, 135–143.

24. Winkler, G. (2001). A stochastic algorithm for maximum likelihood estimation in imaging. *Statistics & Decisions*, **19**, 101–120.

25. Winkler, G. (2003). *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction* (2^{nd} Ed). Springer–Verlag.

26. Younes, L. (1991). Maximum likelihood estimation for gibbsian fields. In *Spatial Statistics and Imaging*, (A. Possolo, Ed.), Institute of Mathematical Statistics. Lecture Notes–Monograph Series, 403–426.