Procesos puntuales como herramienta para el análisis de posibles fuentes de contaminación

J.J. Abellán[a] / **M.A. Martínez-Beneito**[b] / **O. Zurriaga**[b] / **G. Jorques**[b] / **J. Ferrándiz**[c] / **A. López-Quílez**[c]
[a]Servei d'Estadístiques Econòmiques. Institut Valencià d'Estadística. [b]Servei d'Epidemiologia. Direcció General per a la Salut Pública. Conselleria de Sanitat. Generalitat Valenciana. [c]Departament d'Estadística i Investigació Operativa. Universitat de València.

Correspondencia: Juanjo Abellán. Institut Valencià d'Estadística. Servei d'Estadístiques Econòmiques i Treball de Camp. Plaça Nàpols i Sicília, 10. 46003 València.
Correo electrónico: abellan_jua@gva.es

Recibido: 14 de marzo de 2002.
Aceptado: 23 de mayo de 2002.

(Point processes as a tool for analyzing possible sources of contamination)

Resumen

El análisis de un patrón puntual engloba una serie de técnicas que permiten estudiar la distribución de un conjunto de eventos ocurridos sobre una región del plano. Este problema surge en epidemiología cuando se investiga una potencial fuente de contaminación ambiental alrededor de la cual se sospecha que surgen casos de una determinada enfermedad.

En el presente trabajo, se explica brevemente en qué consiste el análisis de un patrón puntual y se ilustra con una aplicación a la determinación del origen medioambiental y al estudio de las zonas de mayor riesgo de incidencia en un brote de neumonía por *Legionella* ocurrido entre mediados de septiembre y principios de octubre en la ciudad de Alcoi (Alicante).

El estudio permitió confirmar el origen medioambiental del brote y señalar las zonas de la ciudad con mayor riesgo, convirtiéndose en el argumento básico para llevar a cabo una exhaustiva inspección de las instalaciones generadoras de aerosoles, tras la cual, hasta la fecha, cesaron los brotes epidémicos.

Palabras clave: Proceso puntual. Análisis espacial. *Legionella*. Contaminación ambiental.

Summary

Point pattern analysis pattern comprises a series of techniques that enables the distribution of a series of events occurring in the vicinity of a particular region of a map to be studied. In epidemiology, this problem arises when a potential source of environmental contamination, possibly leading to cases of a specific disease, is investigated.

The present study provides a brief description of point pattern analysis. The approach is illustrated through determination of the environmental source and study of the areas of greatest risk of incidence of an outbreak of legionella pneumonia that occurred between the middle of September and beginning of October in the city of Alcoi in Alicante (Spain).

Point pattern analysis was able to confirm the environmental source of the outbreak and identify the areas of the city at greatest risk. This provided the justification for an exhaustive inspection of the installations generating aerosols after which, to date, the epidemics ceased.

Key words: Point process. Spatial analysis. *Legionella.* Environmental pollution.

Introducción

El análisis de un patrón puntual, esto es, de la disposición de un conjunto de eventos sobre una región del plano se enmarca en una de las tres grandes ramas de la estadística espacial, la de los procesos puntuales. Básicamente, pretende determinar si dichos eventos presentan un patrón de agregación (los eventos se producen cerca de otros eventos), de inhibición (los eventos aparecen diseminados) o de aleatoriedad espacial completa (los eventos se producen con igual probabilidad en cualquier punto del espacio, con independencia de dónde se hallen otros eventos). Además es posible la comparación de los patrones de dos conjuntos de eventos y, si el patrón es de agregación o de inhibición, puede considerarse su modelización como proceso puntual[1,2], lo que puede permitir análisis estadísticos más ricos.

Ya se han aplicado con éxito en algunas situaciones de interés en epidemiología, tales como el análisis de posibles fuentes de contaminación alrededor de las cuales surgen casos de alguna enfermedad[3-5]. Su campo de aplicación es más amplio, y puede incluso servir también para cartografiar enfermedades, aunque con la ventaja de no trabajar con unidades administrativas.

El presente trabajo pretende introducir las técnicas fundamentales del análisis de patrones puntuales, con una vocación eminentemente práctica, evitando en lo

posible el rigor formal, e ilustrarlos mediante una aplicación a un caso real. En concreto, se ensaya la técnica en el análisis de un brote de neumonía por *Legionella*.

Entre el 16 de septiembre y el 8 de octubre de 2000, se produjeron en la ciudad de Alcoi (Alicante) 57 casos de neumonía por *Legionella*. Suponía la tercera onda epidémica en un año, y fue la más virulenta. De hecho, 40 de los 57 casos iniciaron síntomas en 8 días. Se sospechaba, prácticamente desde el principio, de un origen medioambiental, concretamente de las instalaciones de refrigeración generadoras de aerosoles de empresas y talleres emplazados en el casco urbano. Se habían tomado medidas de desinfección, limpieza e incluso, en algún caso, cierre de dichas instalaciones, mas como se volvía a repetir una nueva onda epidémica, se sospechaba que pudiera haber instalaciones de cuya existencia no tuvieran conocimiento las autoridades municipales y, por tanto, no hubiesen sido sometidas a los controles mencionados para garantizar su seguridad para la salud pública.

Ante esta tesitura, era perentoria la investigación de nuevas instalaciones potencialmente de riesgo y cuya presencia no hubiese sido comunicada a las autoridades tras el reclamo realizado a tal efecto desde el Ayuntamiento, por no sospechar los propietarios su potencial riesgo. Para ello, se pretendía delimitar las zonas de mayor riesgo con objeto de concentrar en ellas los esfuerzos de búsqueda.

A tal fin, se consideró un conjunto de 65 controles que habían formado parte del estudio epidemiológico de la primera onda epidémica. Estos controles se seleccionaron de forma apareada a aquellos casos, con similar edad y con una consulta al hospital una semana antes o después de que el caso ingresara en el mismo. El patrón de los controles se comparó con el de los casos.

Material y métodos

Un patrón puntual está formado por un conjunto de *n* eventos (en nuestro estudio casos incidentes de neumonía por legionella) que han ocurrido sobre unas determinadas coordenadas $\{(x_1, y_1),...,(x_n, y_n)\}$ del plano. En nuestro caso, tanto para el patrón de los 57 casos como para el de los 65 controles, se consideraron las coordenadas de los domicilios de los mismos, que fueron geocodificados a partir de una cartografía de la ciudad a escala 1:2.000 proporcionada por el Ayuntamiento.

A semejanza de la descripción que se realiza de cualquier conjunto de observaciones, mediante una medida de tendencia (media o mediana) y una de dispersión (desviación típica), un patrón puntual se puede describir mediante una medida de tendencia, como la densidad de eventos en la región, y mediante una medida del grado de dispersión o agregación de los mismos. Estas dos características se corresponden con dos funciones, a saber:

1. La función de intensidad[1] $\lambda(x,y)$, que mide la densidad de casos por unidad de área en el punto (x,y). La estimación de la función de intensidad en un punto cualquiera (x,y) de la región en estudio se hace a partir de la posición $\{(x_1,y_1),...,(x_n,y_n)\}$ de los eventos, y puede realizarse de forma puntual o mediante una función auxiliar llamada «núcleo». Esta última forma es la más común, ya que la primera puede considerarse un caso particular suyo. Su expresión es:

$$\hat{\lambda}_r(x,y)=\sum_{i=1}^{n} f_r[(x,y)-(x_i,y_i)]$$

donde $f_r(x,y)$ es la función núcleo[6], que es una función de densidad de probabilidad bivariante simétrica que depende de un cierto parámetro *r*. Este parámetro controla el grado de suavización en el cálculo de la intensidad, de forma que, cuanto más grande sea su valor, mayor será la suavización conseguida, y viceversa. La elección de este parámetro es crucial, y es incluso más importante que la de la función núcleo. A continuación se muestran dos ejemplos de funciones núcleo:

a)

$$f_r(a,b) = \left\{ \begin{array}{ll} \frac{1}{\pi r^2} & \text{si } a^2 + b^2 \leq r^2 \\ 0 & \text{si } a^2 + b^2 > r^2 \end{array} \right\}$$

b)

$$f_r(a,b) = \left\{ \begin{array}{ll} \frac{3}{7} \cdot \frac{1}{\pi r^2} \cdot \left(1 - \frac{\sqrt{a^2 + b^2}}{7r^2}\right) & \text{si } a^2 + b^2 \leq r^2 \\ 0 & \text{si } a^2 + b^2 > r^2 \end{array} \right\}$$

La estimación de la intensidad en un punto, $\hat{\lambda}(x,y)$, proporcionada por la función de la expresión *a)*, sería simplemente el número de eventos ocurridos en el círculo de centro (x,y) y radio *r* dividido por el área del círculo; es decir, la densidad de eventos. En nuestro caso, hemos empleado la función de la expresión *b)*, denominada núcleo cuártico, que presenta ciertas propiedades estadísticas. Se ha tomado $r = 150$ m, siendo la elección de este valor arbitraria, si bien se ha pensado en distancias que pueden tener sentido habida cuenta de la naturaleza del problema. Esto es así porque aunque existen «recetas» para su cálculo[1] e incluso para su posible estimación[6,7], éstas están orientadas a problemas de ecología y sus autores advierten de que pueden no funcionar adecuadamente en otros contextos, lo cual hemos comprobado fehacientemente en nuestro caso.

2. La función $K(s)$ de Ripley[8], que mide la agregación de los eventos a una cierta distancia *s* y que se define según

$$K(s)=\frac{\text{promedio de eventos a distancia} \leq s \text{ de otro evento}}{\lambda}$$

donde el numerador se estima mediante una función tipo núcleo también y el denominador se estima como la densidad media de eventos en la región bajo estudio.

Como se ha mencionado antes, el objetivo es comparar si el patrón puntual definido por los casos y el definido por los controles se asemejan o no, es decir, queremos contrastar la hipótesis:

H_0: casos y controles son dos muestras independientes de la misma población a riesgo.

Este contraste se resuelve a partir de las funciones de agregación de ambos patrones, esto es:

– $K_{controles}(s)$, que representa la agregación espacial propia de la población en riesgo, y que viene dada por la distribución de la población en la ciudad.

– $K_{casos}(s)$, que representa la agregación espacial de la población en riesgo más la agregación adicional provocada por la enfermedad.

Si la enfermedad no induce una agregación adicional, es decir, si verdaderamente los casos no tienden a situarse alrededor de las instalaciones de riesgo, tendría que ocurrir que:

$$K_{casos}(s)=K_{controles}(s)$$

o equivalentemente:

$$D(s)=K_{casos}(s)-K_{controles}(s)=0$$

Por tanto, la hipótesis a contrastar anteriormente citada se puede escribir como

$$H_0: D(s)=0$$

La distribución en el muestreo del estadístico $D(s)$ bajo la hipótesis nula se calcula por simulación. En concreto, se emplea lo que se llama etiquetado aleatorio[9], que consiste en juntar los casos (n_{casos} eventos) y los controles ($n_{controles}$ eventos) y: *a)* poner aleatoriamente a n_{casos} de los $n_{casos} + n_{controles}$ eventos la etiqueta de caso y al resto la de control; *b)* calcular $K_{casos}(s)$ con los eventos etiquetados aleatoriamente como casos y $K_{controles}(s)$ con los etiquetados como controles, para unos ciertos valores de s, $\{s_1,...s_k\}$, y *c)* calcular $D(s_i)=K_{casos}(s_i)-K_{controles}(s_i), \forall i=1,...k$.

Este proceso se repite un número suficientemente grande de veces para obtener muchos valores de $D(s)$ bajo la hipótesis nula. Finalmente, se pueden calcular, a partir de los valores simulados, los percentiles 2,5% y 97,5% que nos darían los límites de un intervalo aproximado de confianza del 95%, y representar finalmente el valor observado del estadístico $D(s)$ con los casos y controles originales frente a los límites de confianza. Si está dentro de dichas bandas de confianza, significa que no hay evidencia contra la hipótesis nula. En cambio, si está fuera de los límites, por encima, quiere decir que hay más agregación en los casos que en los controles, mientras que si, por el contrario, está por debajo de los límites, se ha de interpretar como evidencia de inhibición, es decir, de menor agregación en los casos que en los controles.

Diggle y Chetwynd[10] también proponen calcular el estadístico $D=k\sum_{i=1}^{k}\frac{D(s_i)}{\sqrt{\text{Var}[D(s_i)]}}$ como medida de la diferencia de agregación global entre los patrones, donde Var $[D(s_i)]$ bajo la hipótesis nula también se calcula a partir de las simulaciones del etiquetado aleatorio.

A parte del contraste para saber si hay indicios de mayor agregación de casos que de controles, otra tarea interesante es representar gráficamente alguna medida de riesgo de observar un caso en cualquier punto de la ciudad. Una posible medida de ese riesgo es la diferencia entre la probabilidad observada de encontrar un caso y la esperada. Si tenemos las intensidades de los patrones puntuales, tanto de los casos, $\lambda_{casos}(x,y)$, como la de los controles, $\lambda_{controles}(x,y)$, una estimación de la probabilidad de encontrar un caso en un punto (x,y) genérico es:

$$p(x,y)=\frac{\lambda_{casos}(x,y)}{\lambda_{casos}(x,y)+\lambda_{controles}(x,y)}$$

mientras que la probabilidad esperada de encontrar un caso viene dada, para toda la ciudad, por

$$\frac{n_{casos}}{n_{casos}+n_{controles}}$$

por lo que el riesgo, en un punto dado, $R(x,y)$, se puede expresar como la probabilidad observada menos la esperada:

$$R(x,y)=p(x,y)-\frac{n_{casos}}{n_{casos}+n_{controles}}$$

La estimación de $p(x,y)$ se hace mediante regresión tipo núcleo. A partir de la variable Z asociada a los eventos, definida según

$$Z_i=\left\{\begin{array}{l}1 \text{ si el evento i-ésimo es caso}\\0 \text{ si el evento i-ésimo es control}\end{array}\right\}$$

la expresión del estimador es:

$$\hat{p}_r(x,y)=\frac{\sum_{i=1}^{k} f_t[(x,y)-(x_i,y_i)]\cdot z_i}{\sum_{i=1}^{k} f_t[(x,y)-(x_i,y_i)]}$$

donde $n = n_{casos} + n_{controles}$ es el número total de puntos, y r, como se ha dicho antes, es el parámetro de suavización de la intensidad, de manera que, también aquí, cuanto mayor sea su valor, superior será el grado de suavización de la probabilidad.

Todos los análisis se han realizado con el programa Splus y se han utilizado las funciones de la librería Splancs[11] de libre distribución.

Resultados

El patrón de los controles de la primera onda se comparó con el de los casos actuales en cuanto edad, sexo, actividad laboral (jubilado o no), hábito tabáquico, enfermedad pulmonar crónica y enfermedad renal crónica de base, sin que ninguna de estas variables presentase diferencias significativas entre ambos grupos:

–Edad media, de los casos: 66,84 ± 14,25 años; de los controles: 62,92 ± 16,26. Diferencia no significativa (p = 0,16).

–Sexo: casos, un 57,89% varones; controles, un 69,23% varones. Diferencia no significativa (p = 0,19).

–Jubilados: casos, un 73,68%, controles, un 64,62%. Diferencia no significativa (p = 0,28).

–Fumadores: casos, un 26,32%; controles un, 30,77%. Diferencia no significativa (p = 0,59).

–Insuficiencia renal crónica: casos, un 5,26%; controles, un 4,62%. Diferencia no significativa (p = 0,8).

–Enfermedad pulmonar crónica: casos, un 14,04%; controles, el 16,92%. Diferencia no significativa (p = 0,7).

En la figura 1 se representa la distribución espacial de los casos, controles e instalaciones de riesgo conocidas hasta ese momento que estaban abiertas al menos desde 10 días antes de que el primer caso iniciara síntomas.

Figura 1. Distribución espacial de los casos (círculos), controles (triángulos) e instalaciones generadoras de aerosoles (cuadrados) en la ciudad. Mapa del riesgo de la enfermedad: las zonas sombreadas presentan un riesgo mayor del esperado.

En la figura 2a se pueden observar las funciones de agregación de los casos, $K_{casos}(x,y)$, y de los controles, $K_{controles}(x,y)$. Se aprecia que la agregación espacial de los casos es mayor que la de los controles prácticamente en todas las distancias s consideradas, lo cual se refleja también en el contraste aleatorio que se puede ver en la figura 2b, donde la curva de la función $D(s)$ está completamente fuera de las bandas de confianza del etiquetado aleatorio, especialmente a partir de los 40 o 45 m. El valor del estadístico de la diferencia de agregación global es $D = 463{,}12$ ($p < 0{,}00001$), lo que nos confirma que hay una fortísima evidencia de mayor agregación en los casos que en los controles. Esta concentración «extra» de los casos en el espacio confirma el origen medioambiental del brote.

Figura 2a. Función de agregación K de Ripley para el patrón de los casos y para el de los controles.

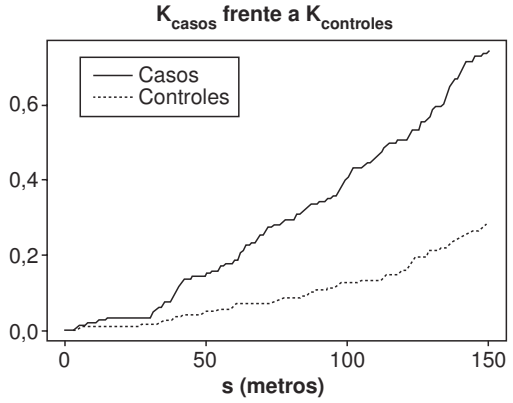
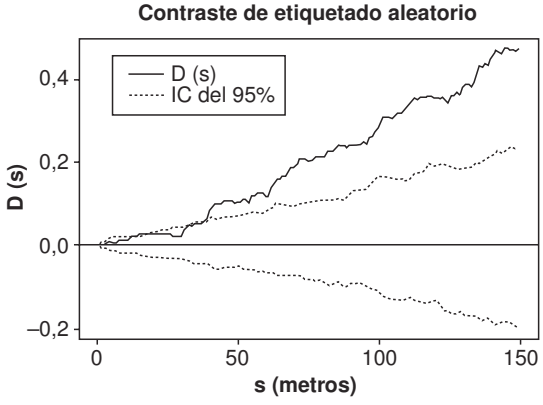


Figura 2b. Diferencia observada de agregación de ambos patrones y bandas de confianza calculadas mediante etiquetado aleatorio.



Finalmente, la probabilidad esperada de que un evento sea caso es 0,463 para toda la región, mientras que la probabilidad observada varía entre 0 y 1, por lo que el riesgo de observar un caso en un punto dado $R(x,y)$ tiene un valor mínimo de –0,463 y un máximo de 0,537. En la figura 1 se presentan sombreadas las zonas donde hay más casos de lo esperado, es decir, donde $R(x,y) > 0$. Para la determinación de las zonas de mayor riesgo se podría considerar, por ejemplo, el cuartil superior la función de riesgo.

Discusión

El parámetro de suavización, *r*, de la intensidad es una de las claves del estudio. Las ideas descritas en la bibliografía para su estimación no funcionaron en nuestro caso, y se probaron unos pocos valores. Valores pequeños (del orden de decenas de metros) producían una focalización excesiva, casi alrededor de cada caso. Valores muy grandes (del orden de centenares de metros) difuminaban las zonas de riesgo en toda la ciudad. Los valores que producían resultados intermedios, más verosímiles, se obtenían con *r* que oscilaban entre 100 y 200 m, por lo que se decidió tomar un valor de *r* igual a 150 m.

El número de controles en el estudio se considera suficiente, dado que la prevalencia de exposición entre los controles se presuponía elevada y esto permitía trabajar con un tamaño de muestra relativamente reducido.

Se codificaron las direcciones de los domicilios de los casos, asumiendo implícitamente que el lugar de residencia era el lugar de exposición para todos los casos, lo que podría no ser demasiado realista. Además, muchos casos eran personas jubiladas que suelen salir a dar un paseo, y ocurría que había un par de rutas muy comunes, por lo que podría haber sucedido que algunas de esas personas inhalaran la bacteria durante esos paseos, aunque únicamente tardaran una o dos horas en realizarlos. Por si esto fuera poco, la bacteria viaja sobre las partículas de vapor de agua, y éste, a su vez, se desplaza con el viento, por lo que la bacteria podría haber llegado a lugares donde no existe ninguna instalación de riesgo.

Sin embargo, a pesar de todas las fuentes de error mencionadas anteriormente, los resultados proporcionados por los procesos puntuales cumplieron ambos objetivos: confirmar el origen medioambiental del brote (dado que los casos presentan una agregación significativamente superior a la de los controles) y revelar las zonas con mayor riesgo de la ciudad, con tres focos repartidos por ellas (dos en el centro y uno en el noreste). La no identificación de las fuentes (no se aisló en la zona de riesgo ni en sus cercanías ninguna instalación generadora de aerosoles con los mismos patrones moleculares de *Legionella pneumophila* aislada en los enfermos) no impidió el control del brote después de la búsqueda exhaustiva de instalaciones de riesgo, control y monitorización de las mismas.

Bibliografía

1. Cressie N. Statistics for spatial data. New York: Wiley, 1993.
2. Diggle P. Statistical analysis of spatial point patterns. London: Academic Press, 1983.
3. Diggle P. A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. JRSS Series A 1990;153:349-62.
4. Diggle P. Point process modelling in environmental epidemiology. En: Barnett V, Turkman K, editors. Statistics for the environment 1. New York: Wiley, 1993.
5. Diggle P, Elliot P. Disease risk near point sources: statistical issues in the analysis of disease risk near point sources using individual or spatially aggregated data. Journal of Epidemiology and Community Health 1995;49:S20-S7.
6. Diggle P. A kernel method for smoothing point process data. Journal of the Royal Statistical Society, Series C 1985;34:138-47.
7. Silverman BD. Density estimation for statistics and data analysis. London: Chapman and Hall, 1986.
8. Ripley B. Spatial statistics. New York: Wiley, 1981.
9. Diggle P, Rowlingson B. A conditional approach to point process modelling of elevated risk. JRSS A, 1994;157:433-40.
10. Diggle P, Chetwynd A. Second-order analysis of spatial clustering for inhomogeneous populations. Biometrics 1991;47: 1155-63.
11. Rowlingson B, Diggle P. Splancs: spatial point pattern analysis code in S-Plus. Technical Report, Lacanster: Lancaster University, 1993.