Detección de cluster espaciales de cáncer pediátrico en los municipios de la Región de Murcia.

López Hernández, Fernando Ant. Fernando.lopez@upct.es
Departamento de Métodos Cuantitativo e Informáticos
Universidad Politécnica de Cartagena

Ortega García. Juan Antonio ortega@pehsu.org
Coordinador PEHSU-Murcia
Hospital Universitario Virgen de la Arrixaca. Murcia.

RESUMEN

El objetivo de este trabajo es la identificación de agrupaciones o cluster de municipios con elevada incidencia de Cáncer Pediátrico en la Región de Murcia. Para alcanzar este objetivo se utilizan diversas técnicas estadísticas englobadas bajo el nombre de Análisis exploratorio de datos espaciales. El principal obstáculo para la obtención de resultados fiables radica en la baja incidencia de la enfermedad que se acentúa con el pequeño nivel de desagregación espacial en el que se analiza la información. Así, en primer lugar se plantea en problema de la inestabilidad de la varianza y con el fin de salvar este obstáculo se plantean diversas alternativas. Los resultados de nuestros análisis muestran en todos los casos la ausencia de spatial cluster (zonas calientes) en la Región de Murcia.

***Palabras claves*:**
Cáncer Pediátrico; Cluster Espaciales;
***Clasificación JEL (Journal Economic Literature)*:** C10, H51

***Área temática*:** Estadística aplicada a los Métodos Cuantitativos

El examen de la ocurrencia de casos de cáncer en la población infantil desde una perspectiva espacial permite a identificar patrones de comportamientos de esta enfermedad. Estos patrones pueden sugerir estudios posteriores más específicos que permitan identificar aquellos factores que inciden en el riesgo de padecer algunos de los tipos de esta enfermedad.
La Región de Murcia esta situada en el sureste de España, bañada por el mar mediterráneo tiene un clima semiárido con inviernos suaves. La población en el año 2.006 era de 1,37 millones de habitantes, de la que el 17,05% era de menores de 14 años. La Región de Murcia tiene una superficie de 11.313 Km^2 y está dividida en 45 municipios, áreas administrativas de tamaño y población muy heterogénea, concentrándose el 52,22% de la población en sólo tres de esos municipios.

Los datos correspondientes a la incidencia de la enfermedad se han obtenido de "Juan Antonio Ortega" y se corresponden con el periodo comprendido entre los años 1998 y 2006, mientras que la información correspondiente a la población procede del Instituto Nacional de Estadística y se refieren a las correspondientes actualizaciones de Padrón Municipal de Habitantes.

Con el fin de evaluar la incidencia del cáncer pediátrico centraremos la discusión en dos indicadores. En primer lugar, y como un indicador bruto, obtendremos la tasa de incidencia (TC *raw rate*) construida como un simple cociente entre el número de casos ocurridos en el periodo analizado y la población de riesgo. Para eliminar la variabilidad de la población a lo largo del periodo analizado (la población de riesgo se ha incrementado en un 14,64% en el periodo analizado y de una forma muy desigual entre municipios), tomaremos como denominador el promedio de la población de riesgo en los años inicial y final del estudio. Así, obtenemos para cada municipio una estimación de la tasa de incidencia, dada por la siguiente expresión

$$\hat{p}_{k,98-06} = \frac{O_{k,98-06}}{[(P_{k,98} + P_{k,06})/2] \times n} \qquad (1)$$

donde k indexa las áreas, n es el número de años, O_k es el número de casos y $P_{k,t}$ la población de riesgo en el año t y área k.

La precisión de esta tasa cruda depende (inversamente) del tamaño de la población y directamente de la proporción desconocida que se desea estimar (p). Por tanto, es difícil comparar estas proporciones entre poblaciones de diferente tamaño como es nuestro caso, en las que las diferentes unidades espaciales (municipios) tienen poblaciones de tamaño muy dispar.

Este problema se conoce como *inestabilidad de la varianza* (Lawson et al. (1999), Waller y Gotway (2004, pp.86–104), Ugarte et al. (2006)) y para solventarlo se recomienda suavizar la tasa cruda mediante una transformación matemática conocida como Empirical Bayes (EB). Este será el segundo indicador con el que trabajemos.

Para detectar la presencia de cluster espaciales con valores similares de incidencia utilizaremos un software específico, GeoDa (version 0.9.5-i) que ya ha sido utilizado con este fin en diferentes trabajos (Rainey et al. 2006, McLaughinn y Boscoe 2007 Kulldorff et al 2006).

Este software permite generar para cada una de las áreas en estudio, un indicador local de correlación espacial conocido como *Local Indicator of Spatial Association* LISA (Anselin 1995) que mide la similitud del valor observado en cada localización con los que se encuentran en su entorno. Así, si el valor de este estadístico para un área determinada es, positivo y significativo, la unidad geográfica puede interpretarse como el centro de un cluster (bien por ser una zona de valores elevados rodeada de valores elevados *High-High* (HH) o por ser una zona de baja incidencia rodeada de zonas con baja incidencia *Low-Low* (LL). Si por el contrario el valor es, negativo y significativo, la unidad espacial puede interpretarse como un atípico espacial, bien por ser una zona de elevada (resp. baja) incidencia rodeada de zonas de baja (resp. alta) incidencia *High-Low* (HL) (resp. *Low-High*(LH)).

2. EL PROBLEMA DE ANALIZAR TASAS EN VARIABLES CON BAJA INCIDENCIA DE CASOS.

Típicamente el *riesgo* de padecer determinada enfermedad, entendiendo como tal la probabilidad de que un determinado suceso ocurra, se estima como el cociente entre el número de individuos padecen dicha enfermedad (O) y el total de *población de riesgo* (P) definida como el número de individuos en los que puede haber ocurrido.

Habitualmente este riesgo está referenciado a un periodo temporal que en el caso de enfermedades de baja incidencia como es nuestro caso suele ser superior a un año.

Así la estimación[1] de la proporción p de individuos que padecen esta enfermedad se denomina *tasa cruda (raw rate)*:

$$\hat{p} = \frac{O}{P} \qquad (2)$$

donde O es el número de individuos que padecen dicha enfermedad dentro de una población de tamaño P. En este caso O se ajunta a una distribución Binimial O=B(P,p) y $\hat{p}$ es el estimador insesgado de p.

Resulta fácil comprobar que

[1] Entendemos aquí "estimación" en un sentido temporal amplio. Efectivamente la información de la que se dispone es la correspondiente a toda la población, pero el periodo temporal de observación es mas o menos pequeño.

$$E[\hat{p}] = p \qquad Var(\hat{p}) = \frac{p(1-p)}{P} \tag{3}$$

En la práctica, esta tasa se suele expresar en número de ocurrencias por 100.000 hab. y cuando el periodo es superior al año (1998-2006 en nuestro caso) se suele hacer la siguiente aproximación:

$$\hat{p}_{98-06} = \frac{O_{98-06}}{[(P_{98} + P_{06})/2] \times 9} \tag{4}$$

Como se observa en (2) la varianza de $\hat{p}$, es decir, la precisión de la tasa cruda depende (inversamente) del tamaño de la población y directamente de la la proporción desconocida que se desea estimar (p). Por tanto, es difícil comparar estas proporciones entre poblaciones de diferente tamaño (como es nuestro caso) en las que las diferentes unidades espaciales (municipios) tienen poblaciones de tamaño muy dispar. Este problema se conoce como *inestabilidad de la varianza*.

De forma mas concreta la varianza de $\hat{p}$ tiene dos problemas. En primer lugar, la proporción desconocida (p) aparece en la expresión de la Var($\hat{p}$), esto se conoce como *dependencia media-varianza*, para valores pequeños de p, el sustraendo en el numerador ($p-p^2$) es prácticamente nulo y por tanto, la varianza es proporcional a la media. Esto se traduce en que aquellas localizaciones con mayor incidencia de la enfermedad presentan mayor grado de inestabilidad.

En segundo lugar, la varianza es inversamente proporcional al tamaño de la población de riesgo, es decir, a menor tamaño de la población (P) la estimación de p mediante $\hat{p}$ es menos precisa. También, cuando las diversas unidades espaciales presentan fuertes variaciones en P la comparación de las proporciones estimadas $\hat{p}$ son mas inexactas. Una importante consecuencia de estas dos cuestiones en lo que respecta al análisis de la información en busca de valores atípicos es que la representación gráfica de la información que sugiera la presencia de "outlier" debe ser tomada con mucha cautela puesto que la presencia de valores extremos puede ser simplemente el resultado del alto grado de variabilidad de la estimación.

La inestabilidad de la varianza de las tasas crudas fruto de la variabilidad de la población entre unidades espaciales ha recibido una extensa atención en el campo del análisis gráfico mediante mapas. Algunas referencias, sin ser exhaustivo: Marshall (1991), Cressie (1992), Gelman and Price (1999), Lawson et al. (1999), Bithell (2000), Lawson (2001b), Lawson and Williams (2001), Lawson et al. (2003), Waller and Gotway (2004, pp.86–104), and Ugarte et al. (2006).

Básicamente los diferentes métodos desarrollados para solucionar el problema pueden dividirse en tres categorías: Transformaciones, Suavización y Regionalización.

El primer grupo de técnicas consiste en transformar la tasa original en una variable diferente con el fin de eliminar la inestabilidad de la varianza (o dependencia entre media y varianza). Una segunda categoría de técnicas, métodos de suavización (smoothing methods)

La tercera categoría de técnicas toma una solución totalmente diferente, e intenta aminorar la inestabilidad incrementando la población de riesgo mediante la agregación de unidades vecinas. Estas técnicas de regionalización ganan en precisión al coste de cambiar la unidad espacial de observación.

La literatura sobre las distintas aproximaciones a la solución del problema es voluminosa y se trata de un área de investigación activa. Algunas recientes comparaciones de las técnicas pueden verse entre otros Kafadar (1994), Gelman et al. (2000), Lawson et al. (2000), and Richardson et al. (2004).

De forma breve y no exhaustiva presentaremos aquí algunas de estas técnicas que posteriormente aplicaremos al caso que nos ocupa.

Rate Transformation

Se han sugerido un buen número de transformaciones con el fin de paliar la inestabilidad de la varianza. El objetivo es obtener otra variable a partir de la original que no sufra estos problemas. El coste asociado al beneficio de la estabilidad de la varianza, es una interpretación mas compleja de la nueva variable.

A simple transformations is the square root transformation, which is easy to implement. More complex transformations include the Freeman-Tukey transformation (Freeman and Tukey 1950), the arcsin (Rao 1973, p. 427), Anscombe (Anscombe 1948), and the empirical Bayes (EB) standardizations (Assun¸c˜ao and Reis 1999). They are briefly considered in turn.

Freeman-Tukey Transformation (FT)

Esta primera opción controla la dependencia entre la media y la varianza, pero no soluciona la desigualdad de la varianza debida a la desigualdad entre las poblaciones en riesgo:

$$Z_i = \sqrt{O_i/P_i} + \sqrt{(O_i+1)/P_i} \qquad (5)$$

donde por el subíndice i hace referencia a la i-ésima unidad espacial de observación. La varianza de esta variable es aproximadamente t2/Pi, donde t2 es una constante que no depende de p. Como se puede observar, la dependencia de P_i en el denominador no se ha eliminado.

Arcsen Standardization

Esta transformación (Ascombe 1948) es

$$X_i = \arcsin\sqrt{\frac{O_i}{P_i}} \qquad (6)$$

La varianza asintótica de esta variable es $1/4P_i$.

Ascombe Standardization

$$X_i = \arcsin\sqrt{\frac{O_i + 3/8}{P_i + 3/4}} \tag{7}$$

***Empirical Bayes Standardization* (EB)**

Recientemente Asunsao and Reis (1999) ha propuesto la EB (Empirical Bayes Standardization) como una forma de corregir el test de Moran de autocorrelación espacial cuando al variable de observación es una proporcion.

Suavización

Las tasas de incidencia pueden suavizarse fácilmente elaborando una media (o mediana) local de tasas ponderadas con o sin ponderar. (see, e.g., Waller and Gotway 2004, pp. 87–88). A continuación consideramos las mas habituales:

Medias locales ponderadas

Este es un ejemplo de medias móviles extendidas al espacio sobre una "ventana". Para cada localización "j" se establece, bajo cierto criterio, aquellas unidades espaciales que se consideran vecinas, J_j y se elabora una media ponderada o media móvil sobre la ventara que se define como:

$$\overline{p}_i = \frac{\sum_j w_{ij}\hat{p}_j}{\sum_j w_{ij}} \tag{8}$$

donde w_{ij} es la ponderación asignada a la relación entre i,j. Tomar ponderaciones que toma el valor 0/1 ignora la diferencia en precisión de los distintos pî, este problema puede eliminarse tomado wij=Pj .

La construcción de la media ponderada depende del criterio de vecindad considerado y de la ponderación asignada.

Medianas locales ponderadas.

Manteniendo la misma filosofía anterior se define

$$\overline{p}_i = \underset{j \in J_i}{mediana}(\hat{p}_j) \tag{9}$$

Este procedimiento puede ser iterado, también puede hacerse ponderaciones.

Regionalización.

El último conjunto de técnicas consideradas para solventar el problema de la inestabilidad de la varianza se centran en el denominador en vez de en el denominador. Puesto que la varianza de la ecuación 4 depende inversamente del tamaño de la población de riesgo, es posible realizar estimaciones estables incrementando la escala espaciadle las unidades de observación. El objetivo de estas técnicas es determinar el área mas pequeña que proporciona estimaciones estables. Este tipo de técnicas se conocen como métodos de *regionalización.*

3. RESULTADOS

En el periodo analizado, desde 1998 hasta 2004, se han diagnosticado 277 casos de cáncer en niños en la Región de Murcia, lo que supone una tasa media de incidencia de 14,1 casos por año para cada 100.000 niños. Como se trata de una enfermedad de muy baja incidencia y el área y la población de estudio es también muy pequeña, las tasas para cada uno de los años de estudio tiene una enorme variabilidad tal y como puede apreciarse en la Tabla 1.

Tabla 1: Tasa anual de incidencia y promedio 9-years por 100.000 niños

Año	Total Casos	Población (0-14 años)	Tasa Cruda
1998	20	203.419	9,83
1999	26	202.701	12,83
2000	31	202.974	15,27
2001	22	204.568	10,75
2002	37	209.976	17,62
2003	44	217.726	20,21
2004	37	222.585	16,62
2005	28	227.773	12,29
2006	32	233.593	13,70

En la Figura 1. se presentan los Box-maps de la TC (Figura 1.1) y de la transformación EB (Figura 1.2). En este gráfico se representan con un mismo color los municipios que se encuentran dentro del mismo cuartil, destacándose en un tono mas oscuro aquellos valores que dentro del tercer cuartil son considerados como valores atípicos.

Figura 1. Box-Map de la TC y de EB

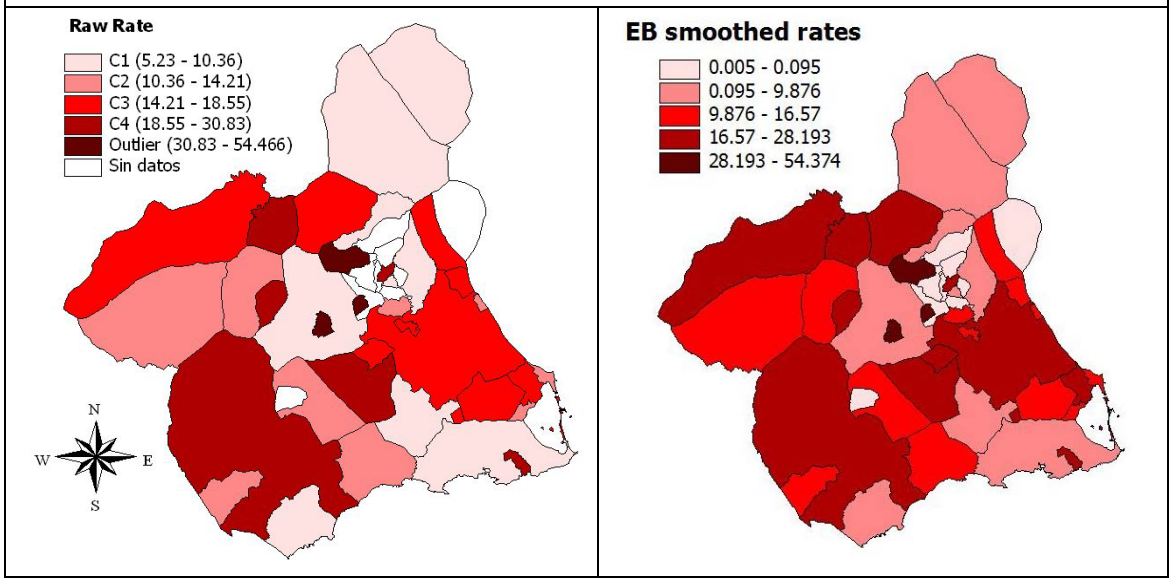


Figura 1.1. Box-maps de la TC (casos por 100.000 niños) por municipios 1998-2006	Figura 1.2. Box-maps de la EB (casos por 100.000 niños) por municipios

La información que suministran ambos mapas es muy similar, con un aspecto ajedrezado donde se alternan las áreas de baja y alta incidencia. Aparentemente no hay indicios de la presencia de ninguna estructura de cluster espaciales que puede refutarse con la obtención de, los índices globales de I de Moran para cada una de estas variables. Ambos índices aceptan la hipótesis de aleatoriedad, descartando por tanto la presencia de una estructura de dependencia espacial.

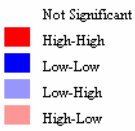
No obstante, debemos resaltar que en ambas variables se detectan tres municipios que presentan valores muy por encima de la media que formalmente son considerados como valores atípicos por exceso que deben ser observadas con cautela.

3.1. Cluster espaciales usando los estadísticos LISA

Un análisis formal sobre presencia de cluster espaciales se obtiene mediante los índices locales de Moran. Los resultados obtenidos para estos índices apenas cambian si utilizamos la TC o su transformación EB, pero si son muy sensibles al criterio que se utilice para asignar vecinos a uno dado. Así, para obtener resultados robustos al criterio de vecindad, hemos utilizado las tres alternativas más comunes.

Los resultados correspondientes a los valores del estadístico para cada uno de estos criterios de vecindad pueden verse en la Figura 3. (Figura 2.1): dos municipios son vecinos si sus capitales se encuentran a menos de 25Km. (Figura 2.2) dos municipios son vecinos si tienen frontera común y (Figura 2.3) se consideran vecinos a uno dado, los 5 mas próximos. Sólo se presentan los correspondientes a la transformación EB porque los de la TC son los mismos.

Figura 2: Municipios con estadístico LISA significativo (p<0.05) para EB rate



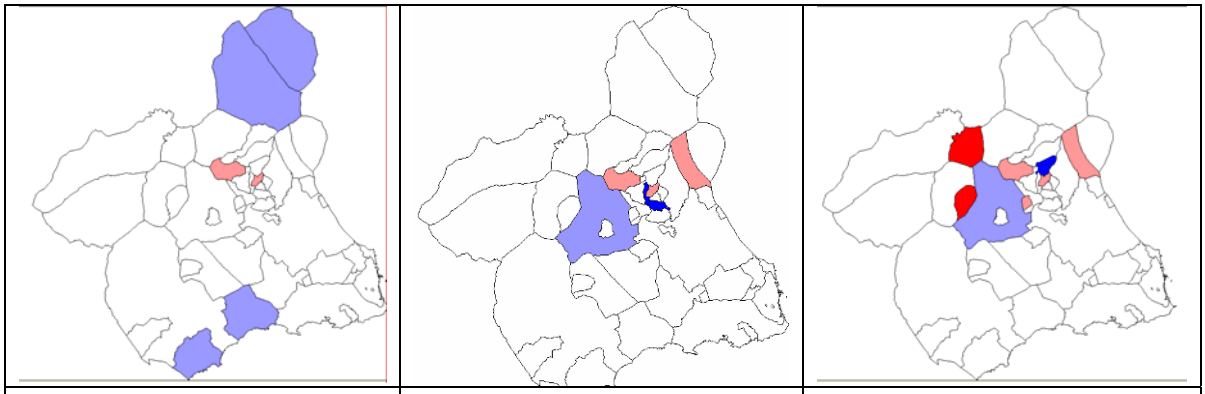


Figura 2.1. | Figura 2.2. | Figura 2.3.

La observación conjunta de los resultados obtenidos en la Figura 2 junto con los de la Figura 1 determina un sistema de vigilancia de incidencia de la enfermedad.

Aunque puede dar la impresión de que la información que suministran estos tres mapas es diferente, éstos mantienen importantes rasgos comunes. Destacaremos de forma detallada las principales resultados.

En primer lugar destacan los municipios de Ricote y Archena en los tres casos aparecen marcados como zonas HL con valores del estadístico negativo y significativo, indicando que se tratan de municipios que presentan tasas de incidencia muy elevadas con respecto a su entorno. El caso de Archena es especialmente llamativo con 7 casos declarados que se corresponde con una tasa de incidencia de 28,19 por cada 100.000 niños y todos los indicios apuntan a que se trata a una zona caliente. En cuanto a Ricote, observando el número de ocurrencias vemos que hay un único caso en todo el periodo y la significatividad del estadístico puede estar afectada por la baja ocurrencia de la enfermdad y debe ser una zona en observación. También el municipio de Fortuna puede encuadrarse dentro de este mismo grupo, ya que en dos de los tres mapas aparece como zona HL pero, al igual que ocurre con Ricote, sólo hay dos casos en los nueve años con lo que el estadístico puede estar *viciado* por la fuerte inestabilidad de la varianza en este caso.

Otras dos zonas calientes son Calasparra (20,63) y Bullas (22,69) que aparecen como centros de cluster del tipo HH cuando se considera el criterio de vecindad (2.3). Además, en el mapa de cuantiles (Figura 1) se encuentran con valores superiores al tercer cuartil, con las tasas de incidencia mas elevadas. Sobre estas dos zonas debe de prestarse una extrema vigilancia en futuros años y deberían intentar encontrar causas…

En el extremo opuesto, zonas frías del tipo Low-High, aparece Mula (8,28) con una tasa de incidencia baja rodeada de valores altos en dos de los tres mapas. También los municipios de Jumilla (5,37) y Yecla (7,57), Mazarrón (10,93) y Águilas (6,25) aparecen como LH en la Figura 3.1

Por último, en la Figura 3.2 aparecen dos municipios marcados como Low-Low. Casi con toda seguridad esto no es mas que el resultado de la estructura espacial de los municipios en esa zona donde hay una estructura espacial de municipios pequeños y bajamente poblados.

4. CONCLUSIONES

Nuestros datos sugieren que no existen cluster espaciales (zonas calientes) aunque hay diferencias en las tasas de SIRs en varios municipios.

Las limitaciones derivadas del estudio por un lado de ser un estudio ecológico y por el otro las dificultades para estudiar una enfermedad como el cáncer pediátrico con baja prevalencia, largos periodos de latencia, y carácter multifactorial con acciones en diferentes periodos críticos (incluso diferentes generaciones). De todo lo anterior, la imposibilidad con la información disponible en los sistemas de registro actuales de obtener datos mínimos y fiables de exposición personal (doméstica y estilos de vida) ó genéticos (como expresión de lo más íntimo en el individuo).

5. REFERENCIAS BIBLIOGRÁFICAS.

- ANSELIN, L., 1995. Local indicators of spatial association – LISA. Geographical Analysis 27 (2), 93–115.
- KULLDORFF (2006) Cancer Map Patterns Am J Prev Med 30 (2S) 37-49
- LAWSON, A., BIGGERI, A., B¨OHNING, D., LESAFFRE, E., VIEL, J.-F., AND BERTOLLINI, R. (1999). Disease Mapping and Risk Assessment for Public Health. John Wiley, Chichester.
- mapping. Statistical Methods in Medical Research, 15:21–35.
- MCLAUGHLIN C. AND BOSCOE F.P (2007) Effect of randomization methods on statistical inference in disease cluster detection. Helath and Place 14 152-163.
- RAINEY, JJ, D. OMENAH, P.O. SUMBA, AM MOORMANN, R. ROCHFORD, AND L. WILSON (2006). Spatial clustering of endemic Burkitt's lymphoma in high-risk regions of Kenya. Inter. J. Cancer, 120 121-127.
- UGARTE, M., IB´A˜NEZ, B., AND MILITINO, A. (2006). Modelling risks in disease

- WALLER, L. A. AND GOTWAY, C. A. (2004). Applied Spatial Statistics for Public Health Data. John Wiley, Hoboken, NJ.