Revista de Matemática: Teoría y Aplicaciones 1999 **6**(2) : 175–187
cimpa – ucr – ccss issn: 1409-2433

UN ENFOQUE DIFERENTE DE LAS TÉCNICAS DE CLUSTERING PARA EL ESTUDIO DE EPIDEMIAS

Gladys Casas Cardoso*– Ricardo Grau Abalo*

Recibido: 10 marzo 1999

Resumen

La clasificación de un brote con la categoría de epidemia requiere del cumplimiento de determinados parámetros epidemiológicos y estadísticos que necesitan de un estudio simultáneo; la teoría matemática ayuda a los epidemiólogos en la detección de las epidemias en aquellos casos en los que no está clara su evidencia. Actualmente estas situaciones se han comenzado a estudiar mediante las denominadas técnicas de *clustering* (del inglés *cluster* que significa aglomeración), apoyados en productos de *software* especializados en el tema. El presente trabajo está encaminado al estudio actualizado de dichas técnicas y a su mejoramiento mediante la inclusión de factores de riesgos. Se expone una aplicación con datos reales.

Palabras-clave: clases de enfermedad, interacción espacio-temporal, estadísticas Scan.

Abstract

Classification of an outbreak with the category of epidemics requires that some epidemiological and statistical parameters, which have to be studied simultaneously, are satisfied; mathematical theory helps epidemiologists in the detection of epidemics in cases when it is not evident. At the present time, these situations are studied with slutering techniques, with the help of specialized software in these topics. The present work aims to study these techniques and its improvement with including risk factors. It is also presented an application with real data.

Keywords: disease clusters, space time interaction, Scan statistics.

A.M.S. Subject Classification: 62H30,68T10,92C60,92T30,92G30.

* Departamento de Matemática, Facultad de Matemática Física y Computación, Universidad Central "Marta Abreu" de las Villas, Carretera a Camajuaní, Km 5 1/2, Santa Clara, Villa Clara, Cuba.

1 Introducción

A pesar de los grandiosos avances en el conocimiento médico, los procesos etiológicos de numerosas enfermedades permanecen aún en la oscuridad. En no pocos de estos casos el saber epidemiológico es esencialmente estadístico [1].

Desde hace varios siglos la teoría matemática, y en particular la estadística, ha dedicado grandes esfuerzos al estudio de las epidemias que tan ferozmente han atacado a la humanidad a lo largo de toda su historia. Se trata de la detección rápida de posibles focos de enfermedades que puedan evolucionar desfavorablemente y caer en situaciones de serias complejidades y de difícil control.

Es por ello que actualmente se le presta especial atención a la determinación de cuando una agregación de casos de enfermos en un área geográfica determinada, en un período de tiempo limitado, o considerando ambos escenarios a la vez, es superior a lo esperado. Si se detecta la presencia de una agrupación de casos de enfermos fuera de lo usual, se puede pensar en el posible origen de una epidemia. En general se hablará en lo sucesivo de técnicas de *clustering* para hacer referencias a problemas de detección de *clusters* y no a su conformación como pretenden las técnicas clásicas y modernas de clasificación jerárquica.

2 Aspectos epidemiológicos en la detección de *clusters* de enfermedades

La modelación matemática o estadística de las epidemias no es un tema novedoso en nuestros tiempos, Farr en 1840 ya daba algunos pasos de avance en este sentido [2]. Sin embargo la mayoría de las investigaciones en este campo se limitaron al estudio de enfermedades transmisibles. La epidemiología en Cuba presta también atención a las enfermedades no transmisibles, pero existe la limitación de que esta teoría de modelación está menos desarrollada. Las técnicas de *clustering* juegan aquí un papel fundamental.

Las agregaciones o conglomerados de enfermos surgen debido a diferentes razones. Se llaman "*clusters* verdaderos" a aquellos que tienen una etiología común o una causa de aparición conocida. Tienen un diagnóstico específico y generalmente se pueden determinar mecanismos de transmisión que los justifican o factores de riesgo asociados a la enfermedad tales como la exposición de los individuos a determinadas radiaciones provenientes de un agente químico, de una Planta Nuclear, o cualquier otra causa común. Algunos de ellos se deben a eventos biológicos [3]. En caso de *clusters* o agrupaciones reales debe ser la cantidad de casos diagnosticados, por sí misma, inusitada y ser estadísticamente significativos; pero también pueden ser significativos *clusters* no verdaderos, algunos de ellos indiscutiblemente se deben a la casualidad. Lamentablemente de todos los casos reportados en la literatura como aglomeraciones, menos del 5% constituyen *clusters* verdaderos [4].

?Cómo distinguir entonces, un *cluster* verdadero de una agregación aleatoria de casos?. No sólo se debe tener en cuenta la significación estadística de los métodos aplicados; deben incorporarse a la investigación todo lo que se conozca de la enfermedad en cuestión, de la población de riesgo y de muchos otros factores que sean relevantes en la comunidad.

Por tanto, no deben utilizarse sólo métodos estadísticos para la detección de *clusters* de enfermedades. En este trabajo se demostrará como pueden enriquecerse las técnicas clásicas de detección de conglomerados en espacio y tiempo con la inclusión de factores de riesgo, lo que ayudará además a caracterizar la enfermedad.

3 Métodos clásicos para la detección de *clusters* de enfermedades

Existe una gran variedad de métodos estadísticos para detectar *clusters* de enfermedades considerando localizaciones geográficas (X, Y), temporales (T) o ambas (X, Y, T).

3.1 Métodos para investigar *clusters* espaciales

Estas técnicas se utilizan cuando existen dudas acerca de cuando una enfermedad es o no más prevalente en localidades o áreas específicas. Pueden estar disponibles dos tipos de datos: tasas dentro de áreas (barrios), o coordenadas de localización geográfica, dada por ejemplo por la dirección de residencia.

En dependencia del tipo de datos disponible se explora cierto patrón espacial. Cuando se dispone de tasas se trata de determinar si las áreas con altas tasas forman *clusters* espaciales, si se dispone de coordenadas se busca determinar cuando localizaciones específicas de casos se agrupan más de lo esperado.

Son muchos los métodos estadísticos para la detección de conglomerados geográficos que se han desarrollado en la última década [1]. Entre ellos se encuentran los tests de Grimson, de Cuzick y Edward, los tests de Moran y el test de Pearson como los más conocidos. El test de Grimson puede usarse con datos en forma de proporciones o de localizaciones, [5]. El test de Cuzick y Edward se usa cuando se dispone de datos con coordenadas [6], mientras que el test de Moran es específico para razones, [7]. El test de Pearson por su parte, trabaja con tasas sobre familias o celdas de tamaño K y es particularmente apropiado cuando se desea verificar la existencia de conglomerados en enfermedades raras, o sea, poco frecuentes, [3].

3.2 Métodos para investigar *clusters* temporales

Cuando se observan casos o proporciones de una enfermedad en intervalos de tiempo consecutivos surge la pregunta sobre en qué medida estos casos se presentan con mayor frecuencia que la esperada producto de una "distribución aleatoria". Las técnicas de *clustering* temporales pretenden dar una respuesta acertada, especialmente cuando los datos disponibles se refieren a series cortas de tiempo, que no pueden ser analizadas usando los métodos convencionales [8]. Se trata, por ejemplo de detectar aglomeraciones inusitadas en el tiempo en una serie que se refiere a celdas diarias, semanales o mensuales.

Aparecen en la literatura más de una docena de métodos que realizan ese tipo de análisis. Entre ellos se encuentran el test de Grimson, test de celdas vacías, prueba de Scan, test de Larsen y test de Dat.

El test de Grimson se usa para identificar un *cluster* entre períodos de tiempo con una elevada proporción de casos. El test de celdas vacías se utiliza para identificar y contar las celdas de tiempo en las cuales no hay casos, como podría esperarse. Es apropiado para enfermedades raras, donde no se esperan casos en demasía en ninguna celda, [9]. El test de Larsen se usa para detectar un *cluster* particular que puede desplegarse a lo largo de varias celdas de tiempo, es útil especialmente para verificar la hipótesis de existencia de un solo *cluster*, [10]. Por el contrario, el test de Dat se usa para verificar la hipótesis de un exceso de casos en uno o más períodos de tiempo que pueden ser o no consecutivos. Esta prueba resulta apropiada para datos correspondientes a series cortas, digamos de 5 o 10 períodos.

Por su parte, el test Scan se usa para identificar un conglomerado de casos en períodos consecutivos de tiempo y es aplicable a la hipótesis de que se desarrolla un *cluster* de casos en dos o más celdas sucesivas. El objetivo es probar la hipótesis nula de que los casos diagnosticados están uniformemente distribuidos contra la alternativa de que existe un *cluster* dentro de algún subintervalo de tiempo [11].

3.3 Métodos para investigar *clusters* en dominios espacio-temporales

El estudio de aglomeraciones espacio-temporales no es la sólo la presencia de *clusters* en espacio y de *clusters* en tiempo. De hecho ambos pueden existir por separado sin que aparezca la interacción. Ella supone que los casos cercanos en el espacio estén además cercanos en el tiempo y por ello este tipo de patrones con interacción es muy útil cuando se desean investigar enfermedades transmisibles. Así, la localización de un evento depende de la localización del evento que lo precede. En las enfermedades infecciosas, para que ocurra la transmisión de una persona enferma a una sana se necesita del contacto personal directo o indirecto a través de vectores. También se pone de manifiesto la interacción, cuando la causa de la enfermedad es la exposición a un agente geográficamente determinado (una sustancia tóxica, ciertos tipos de radiaciones, etc.). Los individuos que estén a la misma distancia del agente, reciben dosis similares en el mismo período de tiempo, y por tanto manifiestan síntomas aproximadamente iguales.

Entre los métodos clásicos está el ya mencionado de Grimson, aplicable también en este caso, el método de Knox, quién partió de examinar no exactamente los intervalos de tiempo entre dos eventos sucesivos, sino más bien de los intervalos de tiempo entre todas las parejas posibles de eventos. En los trabajos referidos estas ideas fueron aplicadas no sólo a distribuciones temporales, sino también a distribuciones espaciales donde la noción de eventos sucesivos no tiene sentido, pero sí toda pareja de eventos geográficos. La extensión de tales ideas al caso de la interacción espacio-tiempo constituyen un hito en toda esta teoría como se muestra por primera vez en [12] y posteriormente en [13], [14].

Se encuentran también entre los más referidos el test de Mantel [15] y el test del “k vecinos más cercanos” [16].

4 Un nuevo enfoque: la integración de los factores de riesgo a la detección de *clusters* de enfermedades

Existen varias razones, teóricas y prácticas, que han motivado la idea de integrar el estudio de factores de riesgo a la detección de conglomerados de enfermos:

1. Las técnicas de *clustering* no son autosuficientes para la detección de epidemias, sino que deben estar integradas a otros métodos epidemiológicos.

2. La necesidad de que dichas técnicas no sólo deben ser capaces de ayudar en la detección de una epidemia, sino también en el esclarecimiento de su naturaleza. Esto es importante en el tratamiento de enfermedades no transmisibles, para las cuales no existe una teoría de modelación desarrollada.

3. El interés de fortalecer los métodos de distinción de *clusters* verdaderos de los *clusters* falsos, que tan frecuentemente aparecen en la práctica.

Todo esto y otras cuestiones más justifican el estudio conjunto de las técnicas de *clustering* y de los factores de riesgo asociados a una enfermedad.

4.1 Algunos conceptos generales

En concordancia con lo anterior se plantea como primera idea la necesidad de estudiar conglomerados no sólo considerando coordenadas espaciales (X, Y), temporales (T) o la interacción de ambas (X, Y, T); sino identificar los conglomerados con más dimensiones: $(X, Y, T,\ F_1,F_2,...,F_n)$ donde los $F_i\ i = 1, 2, \ldots, n$ representan factores en un sentido generalizado.

Entre los F_i se pueden distinguir en particular los factores de riesgos propiamente dichos, esto es, factores de carácter epidemiológico que facilitan o propician la enfermedad. Se llamarán riesgos y en lo sucesivo se denotarán $R_1, R_2, \ldots, R_r$ cuando sea necesario especificarlos.

Se encuentran también entre los F_i otros tipos de factores que representan causas o exposiciones directas a la enfermedad que de una forma mucho más directa la propician; por ejemplo la exposición a un agente tóxico en el caso de enfermedades ocupacionales, o el contacto con vectores en el caso de enfermedades transmitidas por esa vía, entre otras. Se denotarán $C_1,C_2,\ldots,C_c$ cuando sea necesario especificarlos.

Finalmente se incluyen en los F_i un último tipo de "factores", manifestaciones o síntomas de la enfermedad: $S_1, S_2, \ldots, S_s$. Debe recordarse que en la validación de un *cluster* verdadero, en última instancia hay que chequear la coincidencia de síntomas en el conglomerado para tener seguridad de que se trata de una misma enfermedad y no de enfermedades parecidas o de diagnósticos falsos.

4.2 Esbozo de técnicas y una teoría consecuente

Para este tratamiento integrado pueden formularse varios métodos. Los más sencillos consisten en ensayar la combinación de las técnicas clásicas de formación de *clusters* (clasi-

ficación jerárquica), con las técnicas de *clustering* para la detección espacio-temporal de epidemias discutidas en este trabajo.

En efecto, una vez probada la existencia de conglomerados (X, Y, T) de enfermos en una determinada región o período, o considerando ambos a la vez; se puede trabajar con el resto de las dimensiones $(F_1, F_2, \ldots, F_n)$ y formar *clusters* jerárquicos. La repetición de las mismas técnicas de *clustering* espacio-temporales en cada uno de los subgrupos que se forman cuando se consideran los tres tipos de factores explicados anteriormente, debe arrojar información acerca de la verdadera naturaleza de los *clusters* y conocimientos precisos sobre la enfermedad que los provoca.

El principio fundamental que se enuncia es el siguiente: si el *cluster* espacial, temporal o espacio-temporal hallado es verdadero y los factores (riesgos, causas, síntomas) caracterizan bien la enfermedad, entonces la significación de la agrupación deberá desaparecer en subgrupos homogéneos respecto a dichos factores. Por el contrario, si los factores adicionales no caracterizan bien la enfermedad deben mantenerse los conglomerados en forma altamente significativa [17].

Una alternativa de la técnica anterior es la introducción sucesiva de los factores $F_1, F_2, \ldots, F_n$ con la consiguiente segmentación de la población. Si los factores de riesgo están bien identificados (por ejemplo usando la técnica de Mantel-Haenszel [18], análisis discriminante, regresión logística y técnicas de CHAID entre otras), las causas o exposiciones están identificadas o preidentificadas y los síntomas están caracterizados, entonces las sucesivas pruebas de *clustering* en subgrupos con $R_1, R_2, \ldots, R_r$ constantes, en subgrupos con $R_1, R_2, \ldots, R_r$, $C_1, C_2, \ldots, C_c$ constantes y en subgrupos con $R_1, R_2, \ldots, R_r$, $C_1, C_2, \ldots, C_c$, y $S_1, S_2, \ldots, S_s$, constantes, deben tender a disminuir la significación de los aglomerados, y deben además arrojar luz sobre los verdaderos factores de riesgo, causas de la enfermedad y su diagnóstico diferenciado. El mantenimiento de la significación en etapas intermedias es siempre un indicador de que algo nos falta para caracterizar a enfermedad.

Existen otras alternativas para incluir los factores en el estudio mismo de los *clusters*. Se trata de elaborar técnicas específicas para probar las hipótesis de existencia de aglomerados considerando las dimensiones: $(X, Y, F_1, F_2, \ldots, F_n)$, $(T, F_1, F_2, \ldots, F_n)$ y $(X, Y, T, F_1, F_2, \ldots, F_n)$.

Si se trata por ejemplo de un solo factor (que a su vez puede ser resumen en algún sentido de un conjunto de factores y que se haya obtenido como resultado de un análisis discriminante, una regresión logística, o un identificador de grupos después de aplicar un método de formación de *clusters* jerárquicos), el problema se simplifica a estudiar los conglomerados en (X, Y, F), (T, F) o (X, Y, T, F).

El trabajo con factores discretos debe ser objeto de un estudio más cuidadoso, pero el resumen de varios factores (que es lo que nos interesa) se puede siempre buscar como una función continua y el método sería aplicable.

5 Productos de *software* para técnicas de *clustering*

Existen varios paquetes específicos para el análisis de *clusters* de enfermedades, en Cuba se dispone y comienza a usarse el sistema denominado *Cluster*, desarrollado por Tim Aldrich, Wanzer Drane y otros colegas en 1993, el cual consta de 12 técnicas para la detección de conglomerados en espacio, tiempo y en dominios espacio-temporales, [3]. Ya fue lanzado un nuevo paquete, conocido por *Stat! (Statistical software for the clustering of Health Events)*, desarrollado por *BioMedware*. *Stat!* soporta dos tipos de datos básicos: puntos, por ejemplo dirección de residencia de los casos diagnosticados, y áreas, por ejemplo tasas de morbilidad por distrito, [19]. Ninguno de estos paquetes contienen posibilidades de incluir como datos, factores de riesgo o de exposiciones supuestamente comunes a agentes químicos, ni técnicas para analizar conglomerados en estos casos. Como quiera que ese es el objetivo de nuestro trabajo, se ha decidido elaborar un *software* que por el momento permita ensayar y apreciar objetivamente las posibilidades, algoritmos de cálculos y limitaciones de los métodos clásicos, que incluya la mayor parte de los métodos para detectar *clusters*, con recomendaciones para el uso de uno u otro y además las nuevas técnicas de *clustering* con factores de riesgos que se pretenden desarrollar.

Se elaboró entonces la primera versión de un *software* que mejora al que existe actualmente en Cuba, está implementado para ejecutarse sobre Windows y brinda al usuario un ambiente cómodo, [20].

6 Ejemplo de aplicación

6.1 El método Scan para la detección de *clusters* temporales

El método Scan se utiliza para detectar agregaciones de casos de enfermos dentro de períodos de tiempo consecutivos. Un conglomerado temporal puede expandirse a dos o más intervalos [8]. Todos los casos diagnosticados en el área de estudio deben estar ordenados cronológicamente de acuerdo con la fecha de detección de la enfermedad, muerte u otro evento de salud considerado.

Sean $X_1, X_2, \ldots, X_N$ variables aleatorias independientes e idénticamente distribuidas que denotan las fechas de ocurrencias de N eventos en el intervalo $(0, T]$. Se quiere probar la hipótesis nula de que los eventos están uniformemente distribuidos contra la alternativa de que existe un *cluster* dentro de algún subintervalo de $(0, T]$ [11]. Lo primero que hace Scan es definir un intervalo o una ventana de tamaño fijo de acuerdo con la duración esperada de la epidemia. Esto debe hacerse antes de inspeccionar los datos recolectados.

La ventana hallada se desplaza a lo largo de la línea del tiempo y se determinan en cada caso, la cantidad de enfermos asociadas a ella [3].

Sean además:

- t : amplitud de la ventana.
- T : período de tiempo total que se analiza.
- $L = T/t$: fracción que representa el período de tiempo total que se analiza con relación al ancho de la ventana.

- N : cantidad de enfermos diagnosticados en T.
- Λ : número esperado de casos por unidad de tiempo en un proceso de Poisson.
- $n_{y,y+t}$: cantidad de enfermos en la ventana $[y, y+t)$.

Hipotéticamente el estadístico:

$$n = n_t(T) = \max_{0 \le y \le T-t} \{n_{y,y+t}\} \tag{1}$$

es el número de casos que aparecen en una ventana cuando se mueve continuamente a lo largo del tiempo. En la práctica, la ventana $[y, y+t)$ se mueve discretamente a partir de una sucesión de puntos equidistantes $y_1, y_2, \ldots, y_k$ que cubren todo el período de análisis de amplitud T. Se denomina paso del Scan o paso para el desplazamiento a:

$$\Delta y = y_k - y_{k-1} \tag{2}$$

Realmente, el estadístico anterior se estima por su versión discreta:

$$\bar{n} = \bar{n}_t(T) = \max_{1 \le i \le kt} \{n_{y_i, y_{i+t}}\}. \tag{3}$$

La idea del método es que, si existe un conglomerado, el número máximo de casos hallados en la ventana debe ser grande [8]. El test estadístico depende de varios de los parámetros explicados con anterioridad y en esencia calcula la probabilidad p de que aparezcan n o más casos en una ventana. La fórmula que utilizamos para p es la definitivamente propuesta en [21].

$$p = P^*(n, \Lambda L, 1/L) = 1 - Q^*(n, \Lambda L, 1/L) \tag{4}$$

donde $Q*$ puede ser aproximado para cualquier $L > 2$ a partir de sus valores con $L = 2$ y $L = 3$.

$$Q^*(n, \Lambda L, 1/L) \approx Q^*(n, 2\Lambda, 1/2)[Q^*(n, 3\Lambda, 1/3)Q^*(n, 2\Lambda, 1/2]^{L-2} \tag{5}$$

La aproximación (5) es fácilmente calculable usando una microcomputadora personal. El cálculo exacto de $Q^*(n, 2\Lambda, 1/2)$ y $Q^*(n, 3\Lambda, 1/3)$ se basa en el teorema demostrado en [21].

La fórmula (5) puede calcularse también para valores no enteros de L. Esto la diferencia de otras expresiones matemáticas que se usaban con estos fines anteriormente. Además de ser menos restrictiva, Nauss demuestra que esta aproximación es mucho más precisa [21] [22], [23].

La significación p hallada, para un conjunto particular de casos, depende del ancho de la ventana y del paso del Scan seleccionados por el investigador, (por defecto se recomiendan 60 días para el ancho de la ventana y 30, para el paso). Como todo esto son cuestiones subjetivas, resulta de gran utilidad hacer varias repeticiones del método utilizando amplitudes diferentes, especialmente cuando no se conoce la duración esperada de la supuesta epidemia. Además, esta información puede ser muy útil para determinar algunos aspectos importantes de la naturaleza del conglomerado, en caso de existir, y de su etiología [3].

6.2 Análisis de la neuritis. Un estudio integral

Por primera vez se enuncian ideas básicas para integrar el estudio de los factores de riesgo a las técnicas de *clustering*. Se decidió entonces probar dicha unión con los datos reales de la enfermedad más importante ocurrida en Cuba en los últimos años: la neuritis epidémica.

El objetivo es determinar ahora hasta que punto la epidemia que comenzó en 1993 era definible aplicando técnicas de *clustering* con los primeros casos reportados. Se trabajó con datos de la provincia de Villa Clara, Cuba.

Se decidió realizar el estudio utilizando solamente coordenadas temporales, pues en las bases de datos se tenían almacenadas tanto la fecha del diagnóstico como la fecha de primeros síntomas reportadas por los enfermos. Siguiendo criterios de especialistas, se utilizó esta última para la detección de los *clusters*. Se hizo el análisis considerando los primeros 50, 80 y 100 casos reportados y se aplicó el método Scan, se consideraron varios valores para el ancho de la ventana y para el paso del desplazamiento, pues resulta muy difícil, incluso para epidemiólogos expertos en el tema, determinar con seguridad cuales son los apropiados.

A continuación se muestra una de las salidas obtenidas con 50 casos:

EPIDET
Sistema Estadístico para la Detección de Epidemias
12/2/1999 5:03:42 PM

Scan en una Línea

Ancho de la ventana : 60

Paso del Scan : 30

Cantidad máxima de casos hallados en una ventana :26

Valor esperado promedio de casos en una ventana :5.61538462

La significación tiene un valor de : 0.00000004

Los resultados obtenidos se muestran a continuación resumidos en una tabla:

Ancho de la ventana	Paso del desplazamiento	$n = 50$ Significación	$n = 80$ Significación	$n = 100$ Significación
60 días	30 días	0.00000	0.00000	0.00000
30 días	30 días	0.00052	0.00000	0.00000
30 días	15 días	0.00031	0.00000	0.00000

Como se puede apreciar, 50 casos ya son suficientes para detectar un conglomerado temporal. Considerando 80 casos se obtienen resultados altamente significativos, que se mantienen si se continúa aumentando el tamaño de la muestra. Incluyamos ahora sucesivamente un factor de cada tipo, que denominaremos R, C y S respectivamente.

Para evitar los problemas relacionados con la potencia de los criterios y que el volumen de la muestra se vaya reduciendo por las consideraciones sucesivas de los individuos que presentan el factor R, los factores R y C y los factores R, C y S, se decidió ir incrementando la muestra hasta completar 50, 80 o 100 casos para cada uno de los subgrupos.

Entre los factores de riesgo se eligió uno de los de más peso, según un análisis discriminante y de regresión logística previo: el hábito de fumar y la cantidad de cigarrillos

promedio diarios. En este sentido se considera que un sujeto tiene el riesgo R si consume al menos una caja de cigarrillos al día.

Con estos datos se aplicó nuevamente el método Scan para detectar conglomerados temporales. Los resultados se muestran a continuación:

Ancho de la ventana	Paso del desplazamiento	$n = 50$ Significación	$n = 80$ Significación	$n = 100$ Significación
60 días	30 días	0.00012	0.00000	
30 días	30 días	0.02159	0.00070	0.00000
30 días	15 días	0.00392	0.00000	0.00000

En la tabla anterior, para 50 casos se detectan *clusters* temporales, los que se mantienen si se aumenta el tamaño de la muestra. En el método Scan, para 100 casos, la primera celda aparece vacía porque cuando el máximo número de casos hallados en una ventana es grande la aproximación propuesta por Nauss sobreestima la probabilidad y sus valores exceden a 1, por tanto la significación resulta ser negativa. En estos casos deben utilizarse otras aproximaciones, como las discutidas en [22].

Entre las "causas" o "factores de riesgo directos", se consideró la ingestión de productos lácteos y su frecuencia semanal. Se considera que un sujeto tiene este riesgo C si ingiere productos lácteos menos de 2 veces a la semana. En el estudio de riesgo integral este fue el factor más importante relacionado con los déficit nutritivos, pero no el único. Los resultados arrojados por el métodos Scan se muestran a continuación:

Ancho de la ventana	Paso del desplazamiento	$n = 50$ Significación	$n = 80$ Significación	$n = 100$ Significación
60 días	30 días	0.18119	0.00061	
30 días	30 días	0.47123	0.05004	0.00118
30 días	15 días	0.18714	0.00249	0.00000

Se observa una pérdida de significación para los diferentes tamaños de la muestra, que se hace más evidente en los primeros 50 casos.

Durante el proceso de diagnóstico de la enfermedad, cada individuo se clasificó, según sus síntomas en "leve", "moderado" o "severo", S representa entonces la severidad de la enfermedad, o sea, S está presente en aquellos individuos en los cuales la neuritis fue catalogada como severa.

A continuación se muestran los resultados de las aplicaciones del método de Scan sobre los individuos que tienen los factores R, C y S. Del total de personas que padecieron la enfermedad, sólo en 45 aparecieron estos tres factores a la vez, esa es la causa de que no se hagan aplicaciones con diferentes tamaños de muestras como hasta ahora se venía

realizando.

Ancho de la ventana	Paso del desplazamiento	$n = 45$ Significación
60 días	30 días	0.12377
30 días	30 días	0.43216
30 días	15 días	0.14521

En términos generales se mantiene la pérdida de la significación, no tanto si se compara con los resultados anteriores, en los que se consideraron los factores R y C, sino con aquellos en los que sólo se tomó en cuenta R, además de las coordenadas espaciales o temporales.

Debe tenerse en cuenta que para realizar este ejemplo se tomó sólo un factor de cada tipo y no todos los que pudieran aportar alguna información. Si esto se hubiera hecho, los resultados serían mejores. No obstante, nuestro objetivo de confirmar el principio fundamental queda completamente cumplido con dicha selección.

7 Conclusiones

El estudio detallado de las técnicas de *clustering* aplicadas a la detección de brotes epidémicos ha ido evolucionando de forma acelerada en los últimos años, al extremo que hoy se considera una herramienta imprescindible para realizar estas valoraciones. Los resultados aquí expuestos constituyen un estudio ordenado y sistemático de los aspectos de mayor incidencia en el tema. Basado en ello, se enuncian las siguientes conclusiones:

1. Existe una gran variedad de técnicas de *clustering* para la detección y caracterización de epidemias. Su aplicación consecuente exige del trabajo simultáneo con técnicas exploratorias de datos y con métodos de carácter epidemiológico puro.

2. Por primera vez se enuncia y valida satisfactoriamente en la práctica la unión de las técnicas de *clustering* con factores de riesgo, causas y síntomas de la enfermedad y se demuestra la utilidad de este enfoque integrado especialmente en el caso de enfermedades no transmisibles.

References

[1] Marshall, R. (1991) "A review of methods for the statistical patterns of disease", *J. R. Statist. Soc.* **154**: 421–441.

[2] Farr, W. (1840) "Progress of epidemics", *Second Report of the Register General of England and Wales*: 91–98.

[3] Aldrich, T.; Wanzer, D. J. (1993) "*Cluster*", Preprint, The agency for Toxic Substances and Disease Registry Division of Health Studies, U.S.A.

[4] Jacquez, G.; Waller, L.; Grimson, R.; Watenberg, D. (1996) "The analysis of disease clusters, Part I: state of the art", *Infection Control and Hospital Epidemiology* **17**(6): 319–327.

[5] Grimson, R.; Rose, R.D. (1991) "A versatile test for clustering and a proximity analysis of neurons", *Meth. Inform. Med.* **30**: 299–303.

[6] Cuzick J.; Edwards R. (1990) "Spatial clustering for inhomogeneous populations", (with discussions) *J. R. Stat. Soc. (series B)* **52**: 72–104.

[7] Moran, P. (1950) "Notes on continuous stochastic phenomena", *Biometrics* **37**: 17–23.

[8] Jacquez, G.; Waller, L.; Grimson, R.; Watenberg, D. (1996) "The analysis of disease clusters, Part II: introduction to techniques", *Infection Control and Hospital Epidemiology* **17**(6): 385–397.

[9] Grimson, R.; Aldrich, T.; Drane J. (1992) "Clustering in sparse data and an analysis of rhabdomyosarcoma incidence", *Statistics in Medicine* **11**: 761–768.

[10] Larsen, R.; Holmes, C.; Heath, C. (1973) "A statistical test for measuring unimodal clustering, a description of the test and of it's applications of cases of acute leukemia in metropolitan Atlanta, Georgia", *Biometrics* **29**: 301–309.

[11] Nagarwilla, N. (1996) "A Scan statistic with a variable window.", *Statistics in Medicine* **15**: 845–850.

[12] Knox, E. (1964) "The detection of space-time interactions", *Applied Statistics* **13**: 25–30.

[13] Knox, E. (1964) "Epidemiology of childhood leukemia in Northemberland and Durham", *Brit. J. Prev. Soc. Med.* **18**: 17–24.

[14] Knox, E. (1965) "Recognition of outbreaks of acute leukemia and congenital malformations", *In Mathematics and Computer Science in Biology and Medicine*: 227–233.

[15] Mantel, N. (1967), "The detection of disease clustering and a generalized regression approach", *Cancer Research* **27**: 209–220.

[16] Jacquez, G. (1996) "A k-nearest neighbor test for space-time interaction", *Statistics in Medicine* **15**: 1935–1945.

[17] Casas, G.; Grau, R.; Allegret, M. (1997) *"Técnicas de Clustering para el Estudio de Epidemias"*. Tesis de Maestría en Matemática Aplicada, Universidad Central "Marta Abreu" de Las Villas.

[18] Mantel, N.; Heanszel, W. (1959) "Aspectos estadísticos del análisis de datos de estudios retrospectivos de enfermedades." *Journal of National Cancer Institute* **22**(4): 719–747.

[19] Jacquez, G. (1994) *Stat! Statistical Software for the Clustering of Health Events (Software Manual)*. Biomedware, Ann Arbor, MI.

[20] Casas, G.; Grau, R. (1998) *EPIDET: Sistema Estadístico para la Detección de Epidemias por Técnicas de Clustering (Manual de usuario)*. Centro de Estudios de Informática, Universidad Central de Las Villas, Cuba.

[21] Nauss, J. (1982) "Approximations for distributions of Scan statistics", *Journal of the American Statistical Association* **77**(377): 173–183.

[22] Glaz, J. (1993) "Aproximations for the tail probabilities and moments of the Scan statistics", *Statistics in Medicine* **12**: 1845–1852.

[23] Sahu, S.; Bendel, R.; Sison, C. (1993) "Effect of relative risk and cluster configuration on the power of the one-dimensional Scan statistics", *Statistics in Medicine* **12**: 1853–1865.