Análisis de densidad local

Un método alternativo para la selección del radio: el producto cuenta-resolución

Autor:
Díaz, Gonzalo

Tutor:
S. D

1999

Tesis presentada con el fin de cumplimentar con los requisitos finales para la obtención del título Licenciatura de la Facultad de Filosofía y Letras de la Universidad de Buenos Aires en Ciencias Antropológicas

Grado

Gonzalo Díaz

Tesis de Licenciatura en Ciencias Antropológicas

LU: 20.733.261 / 90

Análisis de densidad local

Un método alternativo para la selección del radio:

el producto cuenta-resolución

Introducción

Esta tesis trata sobre un método de análisis de distribución de artefactos llamada "análisis de densidad local". En dicho tipo de análisis, se mantiene como un problema clave la seleccion de un adecuado radio a partir del cual computar las unidades "vecinas" al centro seleccionado. Tradicionalmente, la elección de dicho rango se basa en fines heurísticos a priori, o bien se realiza algún tipo de análisis multivariado sobre las matrices de densidades locales para un rango discreto de radios. Dichos análisis, amén de su carga de presupuestos sobre linearidad y normalidad de las distribuciones, presentan un grado considerable de "oscuridad" matemática que permite su uso (y abuso) por

numerosos autores sin la suficiente aclaración de sus alcances o interpretación de sus resultados.

En contraposición, el método aquí presentado descansa en unas cuantas consideraciones de sentido común, unidas a un tratamiento computacional adecuado pero de escasa complejidad matemática. Además, no incurre en ningún presupuesto sobre la distribución de los artefactos.

El Análisis de Densidad Local

Este análisis fue presentado por Ian Johnson en su tesis en 1976, y referido por el mismo autor con agregados menores en numerosos trabajos posteriores.

El análisis de densidad local (LDA por sus siglas inglesas) es presentado por su autor con las siguientes características diferenciales respecto a otro tipo de métodos:

- es adecuado para colecciones de datos relativamente pequeñas y de alta "resolución"
- es un método descriptivo antes que una modelización estadística
- es matemáticamente simple
- permite realizar inferencias sobre el nivel de clustering de los artefactos y a la vez sobre el nivel de asociación entre categorías de artefactos.
- Ideado para analizar datos bajo la forma de sistema de coordenadas, puede ser fácilmente adaptado para datos en forma de conteo por grillas.

La primera medida de importancia del LDA es la "densidad local".

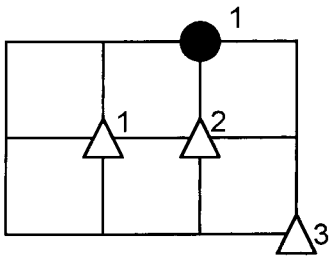
La denisdad local es una relación diádica entre ítems pertenecientes a dos categorías, y puede calcularse para cada uno de los ítems de la categoría "de partida" con relación a la totalidad de los ítems de la categoría "de llegada" (la categoría de partida y la de llegada pueden coincidir, es decir, se puede calcular la densidad local de un ítem de una categoría con respecto a la misma categoría a la que pertenece el ítem).

Llamemos a los ítems de la categoría de partida "centros" y a los ítems de la categoría de llegada "vecinos". Llamemos también "la cuenta de vecinos" o $M_{cv}(r)$ a la cantidad de vecinos (v) que existe alrededor de un centro (c) tomando como límite un determinado radio r.

La densidad local para el k-ésimo ítem de la categoría c está dada enonces por la cuenta de vecinos para ese ítem , dividida la superficie del círculo descripto por el radio. en símbolos

$$\frac{M_{cv}^{k}(r)}{\pi . r^{2}}$$

Para ofrecer un ejemplo numérico, supongamos que en la grilla a continuación existen 2 categorías de ítems: huesos (representados por triángulos) y raspadores (círculos).



Tanto huesos como raspadores tienen sus ítems numerados según un criterio arbitrario. Supongamos el lado de las celdas de 1m. Tomando com radio $r = \sqrt{2}$, obtenemos las siguientes densidades locales:

1. raspadores-huesos:

   1.1. primer ítem de los raspadores respecto a los huesos= $\frac{2}{\pi.\sqrt{2}^2} = 0{,}31$

2. raspadores-raspadores:

   2.1. primer ítem de los raspadores respecto a los raspadores= $\frac{0}{\pi.\sqrt{2}^2} = 0$

3. huesos-huesos:

   3.1. primer ítem de los huesos respecto a los huesos= $\frac{1}{\pi.\sqrt{2}^2} = 0{,}159$

   3.2. segundo ítem de los huesos respecto a los huesos= $\frac{2}{\pi.\sqrt{2}^2} = 0{,}31$

   3.3. tercer ítem de los huesos respecto a los huesos= $\frac{1}{\pi.\sqrt{2}^2} = 0{,}159$

4. huesos-raspadores:

4.1. primer ítem de los huesos respecto a los raspadores= $\frac{1}{\pi.\sqrt{2}^2} = 0{,}159$

4.2. segundo ítem de los huesos respecto a los raspadores= $\frac{1}{\pi.\sqrt{2}^2} = 0{,}159$

4.3. tercer ítem de los huesos respecto a los raspadores= $\frac{0}{\pi.\sqrt{2}^2} = 0$

El siguiente concepto de importancia es la "densidad global" de una categoría. Éste es simplemente el número de ítems de esa categoría dividido por el área de análisis. En nuestro ejemplo, la densidad global de la categoría "raspadores' es $\frac{1}{6m^2}$ y la de la categoría "triágulos" es $\frac{3}{6m^2}$.

Por último, el "índice de asociación " entre dos categorías, dado un radio, es igual a la suma de las densidades locales de cada ítem de la categoría centro, dividida por la cantidad de ítems de la categoría centro (un promedio de densidades locales), todo ello dividido por la densidad global de la categoría vecina. En símbolos:

$$C_{cv(r)} = \frac{\left(\frac{\sum_{k=1}^{N_i} (M_{cv})_k}{\pi . r^2 . N_c} \right)}{\frac{N_v}{A}}$$

A continuación, los índices de asociación para nuestro pequeño ejemplo, para un $r = \sqrt{2}$. Las entradas horizontales representan los centros.

	raspadores	huesos
raspadores	0	0,63
huesos	0,63	0,42

Nótese que la matriz es simétrica, y que la diagonal de asociaciones de cada clase consigo misma no tiene por qué contener valores unitarios.Con el radio también expresado en metros, la unidad es escalar, porque se simplifican entre sí las medidas de longitud.

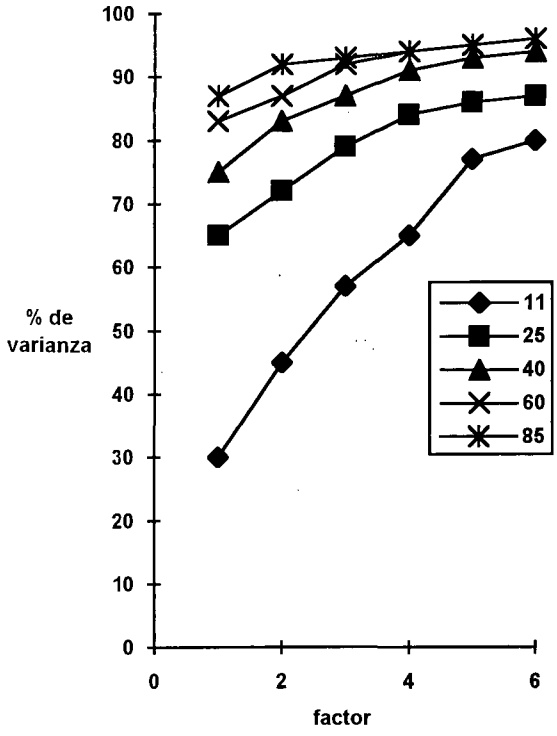
El índice de asociación puede aplicarse tanto a la interpretacion del grado diferencial de asociación entre categorías (a mayor índice, más concentración relativa) o a una única categoría, en cuyo caso una distribución similar a 1 supone una distribución aleatoria de los objetos, un valor superior a 1 una distribución concentrada, y un valor menor que 1 una distribución dispersa.

Obviamente, los valores obtenidos que se vuelquen en la matriz dependerán en gran medida de la selección de un adecuado radio de vecindad. La elección de tal radio es un punto crítico ya que un radio demasiado gande nos dará medidas de asociación altas que harán perder detalle, en tanto que un radio demasiado bajo hará imposible determinar la validez de cualquier interpretación de los pequeños índices de asociación que se obtengan (no se sabrá si se deben a pautas de comportamiento significativas o simplemente a "ruido" en la muestra).

Una primera forma de realizar esta elección de radio es argüir a priori una determinada pauta comportamental (por ejemplo, el área de desecho de determinado material debería tener tal diámetro). Esa posibilidad, al eximir de mayor análisis, no se

trata en este trabajo, aunque algunos autores ponen en duda la validez de este tipo de inferencias (ver Blankholm:1991)

Un camino alternativo sería el tomar una colección de matrices como la obtenida en nuestro ejemplo, pero para distintos radios, y aplicarles algún tipo de análisis multivariado. El autor (Johnson) contemplaba esta posibilidad al momento de elaborar su tesis, pero no disponía aún de herramientas informáticas confiables. Años más tarde (ver Johnson 1984), se publicó un trabajo en el cual se basa en los resultados arrojados por un análisis de correspondencia de las matrices obtenidas para una serie de radios, para determinar el radio óptimo.



El gráfico corresponde a un análisis de correspondencia que el autor realizó sobre los datos del sitio de Princevent, sección 36 V 105. Los ítems fuern agrupados en 24 categorías entre restos esqueletales y artefactos líticos. Los radios de vecindad entregados al análisis fueron 11, 25, 40, 60 y 85 cm. Las matrices de índices de asociación para cada uno de los radios fueron sometidos a un análisis de correspondencia considerando hasta 6 factores.

Nótese que el primer factor ya da cuenta de una gran cantidad de la varianza para los radios mayores que 11. Además, a partir del segundo factor, los radios mayores que 11 muestran un alto nivel de correlación. De todos modos, para evitar el posible efecto de ruido debido a categorías muy dispersas, el autor también descarta el radio 25 cm.

Por lo tanto, el radio seleccionado es el de 40 cm, que explica un porcentaje de la varianza comparable a radios superiores.

La matriz determinada por el radio seleccionado puede someterse, a su vez, a un nuevo análisis de correspondencia para realizar un ploteo en 2 dimensiones. (Aunque todo análisis multivariado se propone a nivel de sugerencia, el LDA en sí mismo termina orgánicamente en las fórmulas de coeficiente de asociación).

Sobre la conveniencia o no de utilizar el método de análisis de correspondencia u otros métodos de análisis multivariado, yo no poseo la sufuciente aptitud matemática para dar un veredicto concluyente. No se trata de ofrecer un método de reemplazo sino alternativo. Los métodos de análisis multivariado y especialmente el análisis de correspondencia gozan de amplia aceptación en no sólo en el análisis de datos de sitios arqueológicos[1] sino para el análisis de distribuciones geográficas en general[2]. Permítasemente no obstante señalar algunas debilidades de este tipo de análisis que son de conocimiento incluso entre sus adherentes.

---

[1] ver Johnson, Whallon, etc
[2] ver Cole & King

Desde un punto de vista general, es aceptado que las técnicas de análisis multivariado (Correspondence analysis, Factor analysis, Principal Component analysis) se emplean a menudo como "caja mágica" en el estudio de campos poco conocidos, en los que intervienen muchas variables empíricas que guardan entre sí relaciones mutuas y complejas. Se trata de reducir las múltiples variables a unos pocos factores de más fácil interpretación, pero cualquier resultado arrojado por un análisis multivariado **no reemplaza**, como parece generalizarse la creencia, a una teoría que dé cuenta de tales factores.

Desde un punto de vista matematico, los cuestionamientos al Correspondence Analysis en particular provienen del llamado "efecto herradura" (en qué medida el segundo eje provee información relevante o representa una distorsión cuadrática respecto al primero, y, en menor medida el tercero representa una distorsión cúbica respecto al segundo, y así sucesivamente)[3].

Por último, desde un punto de vista más subjetivo, la relativa complejidad matemática de estos métodos representa una carga adicional (aunque hoy día relativizada por la abundante oferta de programas de computación que los realizan).

---

[3] cuestionamientos de Gauch, página 93 de Blakholm

Nuevas medidas

Introduzcamos ahora algunas muevas medidas, que ayudarán a la delimitación del radio

índice de vecindad: también con dos subíndices, (centro y vecino) y referida a un radio específico, esta medida varía en un rango de 0 a 1 e indica, tomando a la unidad como 100%, la cantidad de vecinos de ese ítem en el rango de vecindad indicada por el radio, dividida por la cantidad de vecinos con que que efectivamente podría contar si el rango de vecindad abarcara toda el área a estudiar. Aquí se introduce un pequeño refinamiento respecto a la definición de Johnson, ya que el divisor será la cantidad de ítems de la categoría vecína excepto en el caso en que la categoría vecina coincida con la categoría centro. En este último caso al divisor deberá restarsele uno, ya que el centro mismo no entra en la cuenta de vecinos posibles[4].

[4] Johnson no toma esta precaución para el caso de los datos basados en coordenadas de ítems individuales, aunque sí lo hace en la adaptación de su análisis a los datos agrupados por celdas, sin duda porque la omisión de la resta en este caso tiene consecuencias más graves incluso en muestras relativamente grandes.

en símbolos, el índice de vecindad para el k-ésimo ítem de una categoría c, dado el radio k es::

$$V_{cv}(r)_k = \frac{M_{cv}(r)}{d} \quad donde \quad \begin{matrix} d = N_v \Leftrightarrow v \neq c \\ d = N_c - 1 \Leftrightarrow v = c \end{matrix}$$

Promedio de vecindad: el promedio de todos los índices de vecindad de una clase en otra para un determinado radio. en símbolos:

$$P_{cv}(r) = \frac{\sum_{i=1}^{N_c} V_{cv}(r)_i}{N_c}$$

Esta nueva medida también tiene un rango posible entre 0 y 1.

Cuenta ponderada: Esta medida corresponde a un determinado radio, y consiste en un promedio de todos los promedios de vecindad que puedan establecerse provenientes de sendas relciones diádicas posibles.

En símbolos, sea q el número de categorías de una muestra

$$S(r) = \frac{\sum_{i=1}^{q} P_{iq-i}(r)}{q^2}$$

Dado que se establecen relaciones diádicas entre todas las categorías de la muestra, el divisor de esta razón es la cantidad de categorías al cuadrado. Ésto permite que el rango de variación de la cuenta ponderada sea, nuevamente, entre 0 y 1.

Rango de radios de vecindad

Descartado por obvio el criterio comportamental de elegir un radio, único, ensayaremos la forma de elegir un radio adecuado dentro de un radio determinado.

Independientemente de la certeza o no que posean los métodos de Johnsosn para elegir entre una serie de radios propuestos, los criterios para establecer la colección misma de radios a partir de la cual se elige uno brillan por su ausencia. En otras palabras, dada una serie de radios, aún admitiendo que un análisis multivariado de una serie de matrices con los índices de asociación obtenidas a partir de sendos radios podría llegar a indicarnos cuál es el más apropiado, ¿cómo se establece esa serie de radios que originó las matrices?. Obviamente entran en juego una serie de datos implícitos de acuerdo al "ojo" y la experiencia del autor, pero esto resulta inaceptable como método.

Intentaré ir acotando el intervalo de radios posibles que será la fuente de nuestra elección: Una primera acotación resulta obvia: el intervalo irá desde la máxima hasta la mínima distancia cartesiana entre dos ítems cualesquiera.

Los radios se seleccionarán del conjunto de distancias entre los puntos, que es $\binom{N}{2}$. Dependiendo de la potencia del método y las herramientas computacionales que se utilicen, esta medida puede representar problemas. La manera más inmediata de reducir el número de medidas es agruparlas en torno a valors discretos.

Al respecto, se pueden adoptar dos caminos:

- Calcular todas las distancias entre ítems y luego convertirla en un rango discreto de n miembros arbitrarios.
- Hacer discretas las medidas de x e y de las variables (efecto similar a una conversión a grillas) y basarse en esas coordenadas para sacar los radios.

El primer método tiene la ventaja de que el error absoluto es menor, ya que la conversión a discreto se realiza una sola vez. El que las medidas discretas obtenidas no correspondan a distancias reales no representa una desventaja, ya que los radios serán cotas, no hace falta que dos ítems se encuentren precisamente a esa distancia.

La desventaja principal es que requiere el cálculo previo de las $\binom{N}{2}$ distancias.

El segundo método, al reducir la cantidad de coordenadas distintas, reduce también la cantidad de radios. En este caso el error absoluto será mayor, ya que son dos las medidas involucradas en el cálculo de la distancia que se redondean.

Ambos métodos tienen la desventaja de que el número de ítems del rango discreto de radios, o el intervalo entre ellos que se elija para determinarlos, son arbitrarios. (Aunque en todo caso, esta arbitrariedad es menor a la de elegir radios directamente a placer).

Una posibilidad para reducir la arbitrariedad de esta decisión es actuar recursivamente, y el método que propongo permite hacerlo tantas veces como sea necesario:

Un procedimiento posible es:

Determinar un radio exageradamente grande (por ejemplo, igual a la diagonal del área a tratar) y aplicar mi método a un rango pequeño de radios (por ejmplo 2: ese radio máximo y su mitad). Se determina cuál es el radio más adecuado. Si el radio más adecuado es, por ejemplo, el superior, se repite la operación de proponer 2 o más radios,

ahora entre valores que vayan desde la mitad hasta la cota superior. Se determina nuevamente cuál es el más adecuado y así sucesivamente, con la precisión que se desee. La idea es similar a las búsquedas logarítmicas en una lista ordenada.

En última instancia, las operaciones que describo a continuación también se pueden aplicar a un rango arbitrario de radios propuesto por un analista experimentado, para comparar su idoneidad en la selección de uno de esos radios en relación a otros métodos.

Resolución

El concepto de "resolución", aunque intuitivamente simple y fundamental para el tema que nos ocupa, no posee una formulación matemática en el método de Johnson. Daré una definición de esta medida, haciéndola función del radio (una resolución para cada radio posible).

Sea M media matriz cuadrada que tiene por entradas tanto en las filas como en las columnas a todos los pares ordenados posibles enrte categorías. Por ejemplo, si estamos trabajando con 4 categorías la matriz será una matriz cuadrada de 16 X 16. El primer componente representa a la categoría centro y el segundo la vecina. Cada celda

se llena con el valor absoluto de la resta entre el promedio de vecindad entre las categorías indicados por la fila y de las categorías indicado por la columna. La matriz es simétrica y también puede prescindirse de su diagonal que contiene únicamente valores 0 (pues el el valor de ese promedio de vecindad restado a sí mismo).

Elude en este cálculo el usual paso de elevar las diferencias al cuadrado porque no tengo ninguna razón para darle un peso diferencial a las diferencias mayores o menores (recuérdese además que los valores serán todos fraciones menores o iguales a la unidad, y al elevarse al cuadrado darían valores aún más pequeños).

Nótese además que, al tratarse de categorías ponderadas, hay poca probabilidad de que la resolución presente una exagerada sensibilidad a valores grandes y únicos.

Recordemos brevemente que la condicion que debe poseer una medida de radio es, un compromiso: por un lado abarcar un conteo de ítems suficiente como para restarle incidencia al ruido por conteo insuficiente. Por el otro, no acercarse demasiado al máximo posible de ítems abarcados para evitar que el resultado pierda resolución.

Sea q la cantidad de categorías de la muestra

$$R(r) = \sum (P_{ab}(r) - P_{cd}(r))$$

$$a,b,c,d \in N$$
donde $1 \leq a,b,c,d \leq q$.
$$c \prec a \vee d \prec b$$

(El último renglón del "donde" garantiza sólo se seleccionen celdas de la mitad inferior de la matriz). Ésto diferencia la "reslоución' de otras medidas conocidas, como la distancia euclidiana[5].

Empíricamente (y según indica el sentido común), la resolución experimenta un brusco "salto" en los primeros valores de radio, ya que el conteo de ítems "salta" de 0 a algún valor distinto de 0. Luego, a medida que el radio se va aumentando y el conteo de ítems por consiguiente crece, la resolución se va moderando, aunque sería probable que alcanzase aún otros picos si la distribución de una categoría presentara un aspecto visual de "anillos concéntricos" respecto de otra categoría.

Si continúan aumentándose los radios, llegará un momento (como máximo al llegar a la máxima distancia posible entre dos ítems cualesquiera, pero en la práctica mucho antes) en que la resolución se anula totalmente, ya que nuevos aumentos de radio no representan adiciones a los conteos.

---

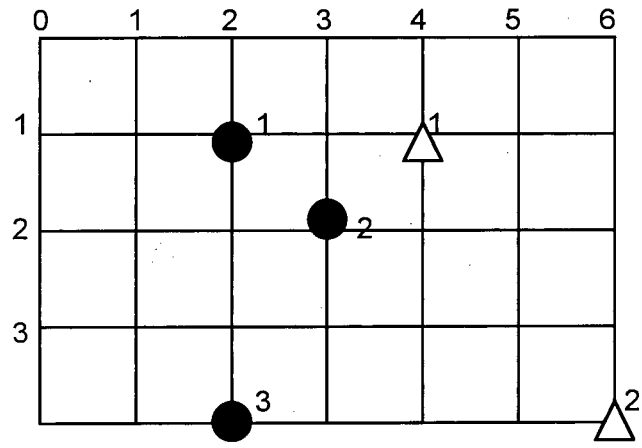
[5] No obstante, la resolución en este tipo de matrices dará siempre un valor igual a ½ de la dstancia euclidiana, aunque ésta resulta computacionalmente más ineficiente ya que calcula el doble de restas, más los resultados "0" de la diagonal.

Los radios de nuestro interés serán aquellos que producen "picos" que no se deban al ruido inicial. Es decir: radios que abarquen un conteo proporcionalmente respetable de ítems, y al mismo tiempo que presenten una resolución comparativamente importante en relación a sus vecinos superiores e inferiores, comportándose como "extremos .relativos" de una gráfica R(r) de resolución en función del radio.

En otras palabras: el radio óptimo deberá presentar la resolución más grande posible pero poseyendo una cuenta ponderada también lo más grande posible. El crecimiento de ambas cantidades guarda una relación inversa (obviamente no lineal), y una operación que representa un compromiso entre ambas cantidades es le producto. Por lo tanto, defino mi última magnitud: el producto cuenta-resolución:

$$X(r) = S(r).R(r)$$

Ejemplifiquemos los cálculos necesarios en una mini-grilla con muy pocos ítems, lo que no permitirá elegir las opciones más exhaustivas posible sin perdernos en una combinatoria demasiado grande.



Coordenadas de los ítems separados por grupos

raspadores	huesos
x=2 ; y=1	x=4 ; y=1
x=3 ; y=2	x=6; y=4
x=2; y=4	

Radios posibles: 1.41 ; 2 ; 2.23 ; 3 ; 3.6 ; 4 ; 5

En las medidas siguientes, la categoría "círculo se representa con una 'c' y la categroría triángulo con una 't' "

Para el radio 1.41

Vcc(1.41)1=0.5	Vct(1.41)1=0	Vtc(1.41)1=0.33	Vtt(1.41)1=0
Vcc(1.41)2=0.5	Vct(1.41)1=0.5	Vtc(1.41)2=0	Vtt(1.41)2=0
Vcc(1.41)3=0	Vct(1.41)3=0		

Pcc(1.41)3=0.33	Pct(1.41)=0.166	Ptc(1.41)=0.166	Ptt(1.41)=0

S(1.41) = (0.33+ 0 + 0.166 + 0.166) / 4 = 0.16

Matriz de restas

	cc	ct	tc	tt
cc				
ct	0.33 - 0.166			
tc	0.33 - 0.166	0.166-0.166		
tt	0.33 - 0	0.166 - 0	0.166-0	

R(1.41)=1

S.R= **0.166**

Para el radio 2

Vcc(2)1=0.5	Vct(2)1=0.5	Vtc(2)1=0.66	Vtt(2)1=0
Vcc(2)2=0.5	Vct(2)1=0.5	Vtc(2)2=0	Vtt(2)2=0
Vcc(2)3=0	Vct(2)1=0		
Pcc(2)=0.33	Pct(2)=0.33	Ptc(2)=0.33	Ptt(2)=0

S(2) = (0.33+ 0.33 + 0.33+ 0) / 4 = 0.25

Matriz de restas

	cc	ct	tc	tt
cc				
ct	0.33 - 0.33			
tc	0.33 - 0.33	0.33 - 0.33		
tt	0.33 - 0	0.33 - 0	0.33-0	

R(2)=1

S.R= **0.25**

Para el radio 2.236

Vcc(2.236)1=0.5	Vct(2.236)1=0.5	Vtc(2.236)1=0.66	Vtt(2.236)1=0
Vcc(2.236)2=1	Vct(2.236)1=0.5	Vtc(2.236)2=0	Vtt(2.236)2=0
Vcc(2.236)3=0.5	Vct(2.236)1=0		
Pcc(2.236)=0.66	Pct(2.236)=0.33	Ptc(2.236)=0.33	Ptt(2.236)=0

S(2.236) = (0.66+ 0.33 + 0.33 + 0) / 4 = 0.33

Matriz de restas de cuentas ponderadas

	cc	ct	tc	tt
cc				
ct	0.66 - 0.33			
tc	0.66 - 0.33	0.33 - 0.33		
tt	0.66 - 0	0.33 - 0	0.33 - 0	

R(2.236)=2

S.R= **0.66**

Para el radio 3

Vcc(3)1=1	Vct(3)1=0.5	Vtc(3)1=0.66	Vtt(3)1=0
Vcc(3)2=1	Vct(3)1=0.5	Vtc(3)2=0	Vtt(3)2=0
Vcc(3)3=1	Vct(3)1=0		
Pcc(3)=1	Pct(3)=0.33	Ptc(3)=0.33	Ptt(3)=0

S(3) = (1+ 0.33+ 0.33 + 0) / 4 = 0.4166

Matriz de restas de cuentas ponderadas

	cc	ct	tc	tt
cc				
ct	1 - 0.33			
tc	1 - 0.33	0.33 - 0.33		
tt	1 - 0	0.33 - 0	0.33 -0	

R(3)=3

S.R= **1.25**

Para el radio 3.6

Vcc(3.6)1=1	Vct(3.6)1=0.5	Vtc(3.6)1=1	Vtt(3.6)1=1
Vcc(3.6)2=1	Vct(3.6)1=0.1	Vtc(3.6)2=0.33	Vtt(3.6)2=1
Vcc(3.6)3=1	Vct(3.6)1=0.5		
Pcc(3.6)=1	Pct(3.6)=0.66	Ptc(3.6)=0.66	Ptt(3.6)=1

S(3.6) = (1+ 0.66 + 0.66 + 1) / 4 = 0.833

Matriz de restas de cuentas ponderadas

	cc	ct	tc	tt
cc				
ct	1 - 0.66			
tc	1 - 0.66	0.66 - 0.66		
tt	1 - 1	0.66 - 1	0.66 -1	

R(3.6)=1.33

S.R= **1.11**

Para el radio 4

Vcc(4)1=1	Vct(4)1=0.5	Vtc(4)1=1	Vtt(4)1=1
Vcc(4)2=1	Vct(4)1=1	Vtc(4)2=0.66	Vtt(4)2=1
Vcc(4)3=1	Vct(4)1=1		
Pcc(4)=1	Pct(4)=0.833	Ptc(4)=0.833	Ptt(4)=1

S(4) = (1+ 0.833 + 0.833+ 1) / 4 = 0.9166

Matriz de restas de cuentas ponderadas

	cc	ct	tc	tt
cc				
ct	1 - 0.833			
tc	1 - 0.833	0.833-0.833		
tt	1 - 1	0.833 - 1	0.833-1	

R(4)=0.66

S.R= **0.611**

Para el radio 5

Vcc(5)1=1	Vct(5)1=1	Vtc(5)1=1	Vtt(5)1=1
Vcc(5)2=1	Vct(5)1=1	Vtc(5)2=1	Vtt(5)2=1
Vcc(5)3=1	Vct(5)1=1		
Pcc(5)=1	Pct(5)=1	Ptc(5)=1	Ptt(5)=1

S(5) = (1+ 1 + 1+ 1) / 5 = 1

Matriz de restas de cuentas ponderadas

	cc	ct	tc	tt
cc				
ct	1 - 1			
tc	1 - 1	1 - 1		
tt	1 - 1	1 - 1	1 - 1	

R(5)=0

S.R= **0**

De los cálculos anteriores se desprende qe el radio de vecindad r=3 es no sólo el que presenta mejor resolución, sino el más alto producto cuenta-resolución. Por lo tanto será el radio elegido.

LDA para r=3

superficie del círculo= 28.274 unidades
área total = 24 unidades

densidades locales

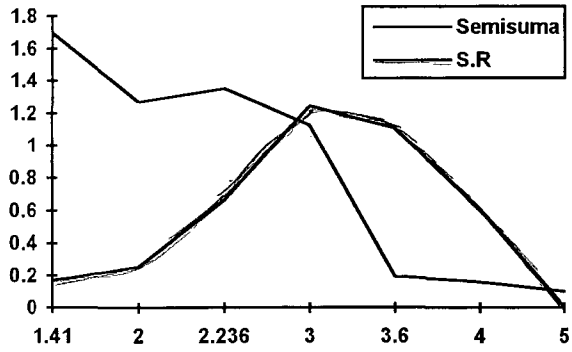
raspadores-raspadores= $\frac{6/3 \cdot 28.27}{3/24}$ =0.565 raspadores-huesos= $\frac{2/3 \cdot 28.27}{2/24}$ =0.282

huesos-huesos= $\frac{0/2 \cdot 28.27}{2/24}$ =0 huesos -raspadores= $\frac{2/2 \cdot 28.27}{3/24}$ =0.282

Tabla de índices de asociación para todos los radios
(el valor del coeficiente de asociación huesos-raspadores se obvia porque la matriz es simétrica).

Se adjunta una columna con la resolución (semisuma de las distancias distancia euclidianas) de la matriz cd coeficientes de asociación para cada radio.
El objetivo es probar que el producto SR no es una mera translation de distancia euclidiana al análisis de densidad local.

radio	Crr	Crh	Chh	Semisuma	S.R
1.41	0.848	0.636	0	1,696	0.166
2	0.424	0.636	0	1.272	0.255
2.23	0.679	0.509	0	1.355	0.666
3	**0.565**	**0.282**	**0**	**1.13**	**1.250**
3.6	0.391	0.391	0.293	0.196	1.111
4	0.318	0.397	0.238	0.16	0.611
5	0.203	0.305	0.152	0.102	0



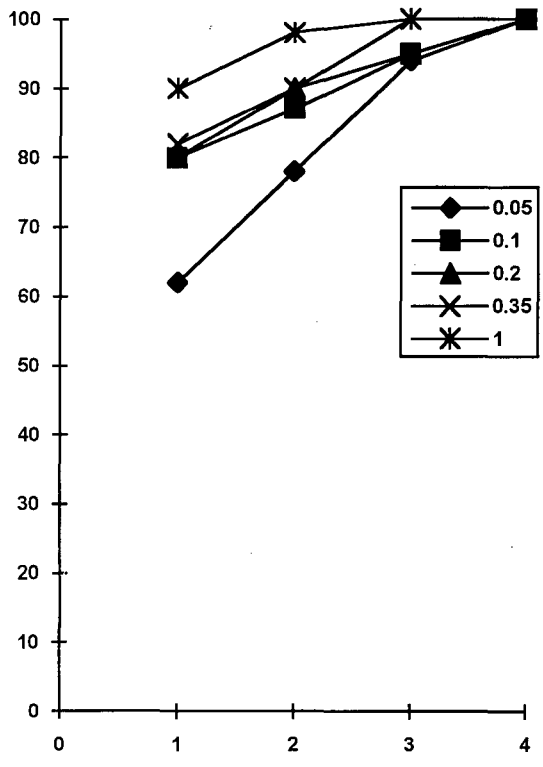
La sim[ple semisuma de dsitancias euclidianas de índices de asociación para r=3 no aparenta tener ninguna relevancia especial respecto a las de los radios vecinos.
En tanto que la semisuma va reduciéndose a medida que se aumenta el radio (resultado de la pérdida de resolución), el prosucto S.R alcanza su punto máximo en r=3.

El Producto SR como criterio de selección aplicado a un rango arbitrario de radios

El PSR puede constituir no sólo un método de determinación del radio adecuado, sino una medida de la idoneidad (en sus propios términos) de un rango discreto de radios arbitrarios, no necesariamente de intervalos regulares.

Como la medida del PSR para cada radio es independiente en sí misma, es decir, no involucra varianzas ni porcentajes con respecto a radios menores o mayores, la aplicación de esta medida presenta una ventaja adicional: su tendencia a crecer o decrecer puede no solamente indicar el radio adecuado, sino sugerir la dirección de un nuevo rango de análisis (por ejemplo en un procedimiento "logarítmico como el mencionado anteriormente)..

Ejemplifiquemos con los datos extraídos del sitio de Mask (Binford: 1978) . Los ítems de este sitio fueron rotulados según sus coordenadas, y agrupados en 5 categorías: artefactos, astillas de hueso, cartuchos, cortezas y huesos largos. Blankholm realizó el cálculo de los índices de asociación para 5 radios arbitrarios, y sometió a las matrices resultantes a un análisis multivariado (Blankholm: 1991). Tras este análisis (Correspondence Analysis), el radio r=0.10 se había revelado el más idóneo.



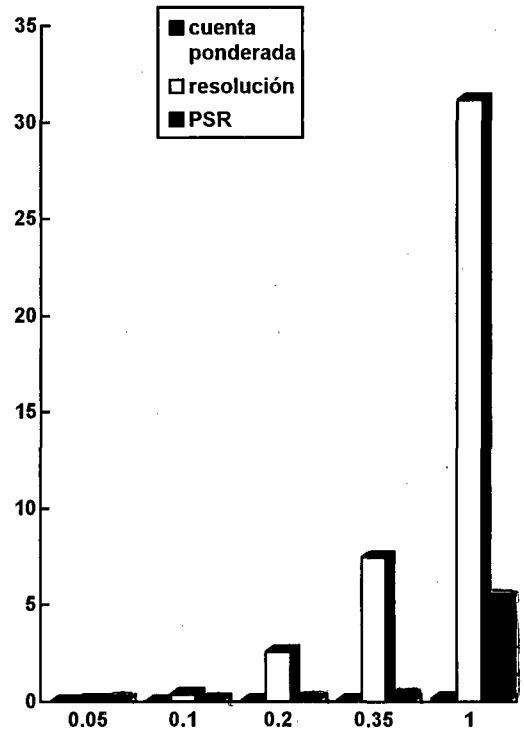
De manera similar al variograma anterior, para determinada longitud de radio el porcentaje de varianza se homologa y pasa a ser es más o menos similar para todos los valores de radio en los sucesivos factores. La estrategia consiste en seleccionar el radio más pequeño de ese "racimo" a partir del cual el aumento de varianza ya no es significativo. Seleccionar un radio mayor provocaría que la relación entre porcentaje de varianza y pequeñez del radio se deteriorase.

El método es conceptualmente correcto, lo arbitrario consiste, una vez más, en la elección inicial del rango de radios al que se aplica. El LDA es un método de análisis post-facto antes que de prospección, por lo tanto no existen constricciones prácticas a priori para ceñirse a un rango de radios determinado. La técnica, utilizada en los dos variogramas que vimos, de ir variando la longitud de entre radios de manera creciente no parece arrojar resultados satisfactorios.

A continuación presentamos la tabla para el producto SR de dichos radios:

radio	cuenta ponderada	resolución	PSR
0.05	0.0002	0.048	0.00001
0.1	0.0017	0.372	0.00063
0.2	0.0109	2.54	0.02781
0.35	0.0331	7.444	0.2469
1	**0.167**	**31.17**	**5.2357**

La medida seleccionada es, obviamente, el radio mayor (eleccion muy distinta al radio de 0.10 que proponía el análisis multivariado). Pero la observación de los valores obtenidos para los demás radios nos proporcionan aún más información: la tendencia enteramente creciente tanto de la resolución como de la cuenta ponderada, y el brusco incremento que experimentan tods las magnitudes en el último intervalo sugiere investigar nuevos radios por encima de r=1m.



Alentados por la anteror tendencia creciente, aumentaremos el intervalo para abarcar un rango de radios hasta el radio máximo. El máximo radio posible puede determinarse mediante métodos computacionalmente sencillos sin necesidad de calcular previamente la totalidad de los radios: se recorre los valores de 'x' y de 'y' buscando el máximo v el mínmo valor para ambas coordenadas, y luego se traza un radio desde el punto de coordenadas mínimas hasta el punto de coordenadas máximas.

x máxima=15.22 y máxima=11.77
x mínima=3.66 y mínima=3.1

El máximo radio posible es, pues, $\sqrt{(15.22-3.66)^2+(11.77-3.1)^2} = 14.45$
Inspeccionaremos radios de 2 en 2 desde 2 hasta 14. Estas medidas máximas y mínimas también nos permiten inferir el área de anásis, que redondeamos en

16mX12m=192m^2. Dicha áres no es necesaria para calcular los valores del PSR, pero sí para calcular los coeficientes de asociación una vez elegido el radio[6].

La tabla a continuación muestra los valores de cuenta ponderada, resolución y producto cuenta-resolución para el rango elegido.

radio	cuenta ponderada	resolución	PSR
2	0.324	110.20	35.72
4	**0.633**	**85.42**	**54.12**
6	0.884	50.95	45.04
8	0.966	18.97	18.33
10	0.995	2.4	2.39
12	1	0	0
14	1	0	0

En la tabla se observa que r=2 (de acuerdo con lo que podíamos esperar del análisis anterior) presenta una enorme resolución. No obstante, no es el valor elegido porque no cuenta con un número proporcionalmente suficiente de ítems, lo que disminuye su PSR. De ahí que el valor elegido es r=4.

[6] En algunos textos (ver Blakholm, Johnson) se dice que el LDA prescinde del área. Esto no es completamente correcto. En realidad, el valor del área total de la zona relevada es necesario para calcular los coeficientes de asociación. En una etapa posterior sel análisis, al emplear las matrices de asociación en un análisis multivariado, el área vuelve a aparecer y puede simplificarse de las cuentas, pero ésto no quita que para calcular los índices de asociación individuales no pueda omitirse el área. De todas maneras, la elección de un área distinta de un rectángulo con las máximas coordenadas está condicionada por factores aún subjetivos, como por ejemplo, si la superficie contorneada está "bien" o "mal" definida por la presencia de puntos.

El PSR y la apreciación de concentración-dispersión.

Vimos anteriormente que la elección del radio a partir del PSR no garantiza una máxima distancia euclidiana en la matriz de índices de asociación. Tampoco el radio elegido satisface la condición por cierto deseable de provocar una provocar una matriz de coeficientes que oscilen entre menores y mayores que uno. De esta manera, y según se explicaba al principio de este trabajo, para esa medida de radio tendríamos una apreciación inmediata del nivel de concentración o dispersión de las categorías entre sí. Esta carencia se puede corregir aplicando nuevamente un simple análisis de correlación (Por ejemplo, Pearson) a las matrices obtenidas, lo cial nos daría matrives simétricas con unos en su diagonal, pero los números obtenidos ya no representan propiamente un LDA.

Véase la presente matriz de coeficientes de asociación para el radio seleccionado r=4 (Las categorías son: artifacts, spent cartridges, wood shavings, bone splinters, large bones)

	artifacts	sp. cart.	w.shav.	b.splint.	large bn.
artifacts	1.42				
sp. cart.	1.9	3.34			
w.shav.	2.04	3.3	2.93		
b.splint.	2.04	3.24	3.18	3.26	
large bn.	1.83	1.36	1.96	2.49	2.57

Aunque se puede extraer conclusiones del valor comparativo entre categorías, todos los valores son superiores a uno.
Compárese con la matriz para el radio seleccionado en primera instancia r=0.1

	artifacts	sp. cart.	w.shav.	b.splint.	large bn.
artifacts	4.88				
sp. cart.	0	44.35			
w.shav.	0	14.20	0.61		
b.splint.	2.82	1.13	2.47	32.54	
large bn.	4.91	1.26	0.61	2.97	19.78

Este radio r=0.1 completa mejor el requisito de oscilar en torno a 1. La mayoría de los coeficientes son superiores a 1, algunos mostrando intensa concentración (madera descortezada vs astillas de hueso, cartuchos vs ellos mismos y madera descortezada). También aparecen índices inferiores a la unidad (artefactos vs madera descortezada y vs catuchos, madera descortezada vs huesos largos). Nótese, no obstante, que los dos primeros son absolutamente 0, situación que cambia para otros radios mayores (ver tablas a continuación) que también poseen índices inferiores a la unidad .

r=0.5

	artifacts	sp. cart.	w.shav.	b.splint.	large bn.
artifacts	9.58				
sp. cart.	0.17	44.33			
w.shav.	0	19.42	83.36		
b.splint.	6.39	1.67	5.27	28.40	
large bn.	3.42	1.52	0.96	3.58	13.17

r=1

	artifacts	sp. cart.	w.shav.	b.splint.	large bn.
artifacts	4.40				
sp. cart.	0.76	25.72			
w.shav.	0.25	18.07	42.58		
b.splint.	5.63	2.05	8.38	18.98	
large bn.	2.59	1.50	0.86	4.07	8.5

Si lo que se buscara fuera una matriz de correlaciones en la cual, además de presentar tanto índices de concentración como de dispersión, todos los índices contaran con cifras significativas, recién un radio de 1m sería el adecuado (pues empezamos a contar con madera descortezada vs artefactos como distinto de 0).

De lo recientemente expuesto concluyo que

La existencia en una matriz de valores que oscilen en torno a 0 (mostrando índices diferenciados en torno al valor límite de concentración y dispersión) dependerá no sólo del método de selección del radio sino de la distribución misma de las categorías de artefactos. Éstos pueden estar simplemente muy concentrados tanto entre sí como con las demás categorías sin importar cuán pequeño o grande fuera el radio de análisis que se elija.
En todo caso, la utilización de un rango arbitrario sometido a un análisis multivariado (es decir, el método presentado en primera instancia) tampoco garantizará obtener esta característica. La existencia de valores superiores e inferiores a 1 es una característica de cada radio individual seleccionado, y no es función de su varianza con respecto a otros radios, que es lo que se dilucida en un análisis multivariado.

Conclusión

El PSR, inicialmente concebido como un método para la selección de un radio adecuado para el Análisis de densidad Local, puede constituir un modesto método en sí mismo para delimitar el radio de áreas circulares de análisis.

Es sensible a la cantidad proporcional de ítems contabilizados, permitiendo desechar situaciones de alta resolución por escaso conteo.

Al constituir cada PSR una medida independiente y no calculada bajo porcentajes o presunciones de normalidad en la distribución con respecto a otras medidas, el PSR puede ser utilizado tanto como método de *determinación* como de *elección* de un radio (siempre sobre datos a priori, no es un método de prospección). En ambos casos, se contará además con la ventaja de poder ir realizando refinamientos sucesivos al rango de análisis elegido, con el detalle que se desee.

Su falla principal consiste en no presentar una lectura diferenciada inmediata de las medidas de concentración y dispersión entre categorías, sacrificando esta característica en función de un conteo más representativo del total de la muestra.

Bibliografía

Blankholm, H.P (1991) *Intrasite Spatial Analysis in theory and practice.* Aarhus University Press
Davis, J (1986) *Statistics and Data Analysis in Geology* John Wiley & Sons
Johnson, Ian
(1976) "Contribution methodologique a l'étude de la repartition des vestiges dans des niveaux archeologiques", his doctorate thesis, presented at the Institut de Quaternaire, Université de Bordeaux.
(1984) "Cell frequency recording and analysis of artifact distributions" in H.Hietala (ed.) *Intrasite spatial analysis in archaelogy*, pp 75-96, Cambridge CUP.
Gauch, H.G. (1982) *Multivariate analysis in community ecology.* Cambridge, CUP.
Hill, M.O. (1974) "Correspondence Analysis: a neglected Multivariate Method. " Journal of he Royal Statistical Society, Series C, 23: 340-354
Hill, M.O. and H. G. Gauch: "Detrended Correspondence Analysis: an improved Ordination Technique". Vegetatio 42:47-58.