APLICACIÓN DEL ÍNDICE RAND PARA MEDIR LA PRESERVACIÓN DE LA ESTRUCTURA ORIGINAL DE DATOS EDÁFICO-TAXONÓMICOS EN UNA PARTICIÓN OBTENIDA CON LA TÉCNICA DE *CLUSTER ANALYSIS*

Jorge Rodríguez Gómez*
Luisa Fernández De Andrade**

RESUMEN

En un estudio realizado por Fernández y Rodríguez (1995), se requirió de la aplicación de la técnica de Análisis de Conglomerados (Cluster Analysis) para determinar, matemática o automáticamente, la clasificación en 7 clases de 30 perfiles de suelos venezolanos (8 Inceptisols, 12 Ultisols, 5 Aridisols, 2 Oxisols, 1 Entisol, 1 Vertisol y 1 Alfisol) formados bajo distintos ambientes naturales. Para medir la calidad de la partición obtenida se decidió utilizar el índice o estadístico Rand, sugerido por Milligan (1980) como un estadístico apropiado. El mencionado índice se determinó con base a una matriz cuadrada y simétrica de orden 30, en donde sólo se compilaban como valores posibles de los elementos de la matriz a la unidad y el cero. El valor del mencionado índice, utilizando la parte triangular inferior de la matriz simétrica, fue de 0,77; cifra que sugiere que la clasificación automática lograda conserva en un 77% la estructura original que consistió en la clasificación de los 30 perfiles en taxones a nivel de orden. Esta tasa de recuperación se considera pertinente si se toma en cuenta que el valor máximo posible de recuperación es del 100%. El propósito de este artículo es, básicamente, mostrar algunas características operacionales del mencionado estadístico así como sugerir su uso a los geógrafos interesados en los tópicos de clasificación de sus objetos de estudio.

Palabras clave: Cluster Analysis, Indice Rand, Taxonomía de Suelos, Análisis de Particiones, Matriz Matemática.

* Profesor Asociado de la Universidad Central de Venezuela, Facultad de Humanidades, Escuela de Geografía, Los Chaguaramos 1041, Caracas, Venezuela.
** Profesora Asistente de la Universidad Central de Venezuela, Facultad de Humanidades, Escuela de Geografía, Los Chaguaramos 1041, Caracas, Venezuela.

ABSTRACT

In a study carried out by Fernández y Rodríguez (1995), application of cluster analysis was required to determine mathematically or automatically, the classification in 7 classes of 30 profiles of Venezuelan soils (8 Inceptisols, 12 Ultisols, 5 Aridisols, 2 Oxisols, 1 Entisol, 1 Vertisol y 1 Alfisol) formed under various differents natural environments. In order to measure the quality of the obtained partition to use the Rand Index was considered, as suggested by Milligan (1980), as a suitable statistic. This index was determined on a square and symmetric matrix basis of order 30, where only possible values of matrix elements to unity and zero were compiled. The value of such index, using the lower triangular part of the symmetric matrix, was 0,77; value which suggests the automated classification obtained preserves 77% of the original structure which consisted in the classification of the 30 profiles in taxons at level order. This rate of recovery is considered appropriate if it is taken in account that the maximum possible value of recovery is 100%. The purpose of this article is to show basically some operational characteristics of the Rand Index, and also to suggest its use to geographers interested in classification topics.

Key words: Cluster Analysis, Rand Index, Soil Taxonomy, Partition Analysis, Mathematical Matrix.

****

INTRODUCCIÓN

Bajo el nombre general de *Cluster Analysis*, que ha sido tradicionalmente traducido como *Análisis de Conglomerados* y los autores se permitirán denominar, equivalentemente, como *Análisis de Particiones*, se abarcan un grupo de estrategias cuantitativas cuyo propósito es identificar, en un conjunto dado de casos o *individuos*, los subconjuntos que son similares en relación con diversas variables o atributos seleccionados en tales individuos.

La *partición simple* obtenida (dos o más subconjuntos no yuxtapuestos de casos similares dentro de un conjunto mayor) es una agrupación *objetiva* de considerar al conjunto de individuos, agrupación que frecuentemente no coincide con la estructura original, previamente conocida, que constituían dichos casos antes de someterla al tratamiento clasificatorio matemático. En ese sentido, se considera que se ha logrado

una partición pertinente si ésta ha preservado, en una alta proporción, la estructura original. La figura 1 intenta esquematizar las fases que constituyen, de manera general, la matematización del proceso de clasificación.

En ese orden de ideas, este artículo tiene como objetivo básico mostrar cómo, mediante el índice o estadístico Rand, se evalúa en qué medida la *clasificación* obtenida mediante la técnica del Análisis de Particiones (que abreviadamente se resumirá con el acrónimo AP) conserva o preserva la estructura original de los datos, previamente determinada según criterios *racionales o lógico conceptuales* de un área de conocimiento científico en particular.

ANTECEDENTES

Milligan (1980) publica una investigación donde muestra cómo puede evaluarse la preservación o recuperación de un conjunto estructurado de datos a los cuales subsiguientemente se le ha aplicado alguna *estrategia* matemática de agrupamiento. En ese sentido, Milligan (op.cit.; p.330) considera que hay, al menos, dos criterios de evaluación: (i) externo e (ii) interno. Un *criterio externo* de evaluación es el que mide la preservación original de la estructura con base a información independiente o externa al proceso matemático de agrupamiento; por otra parte, el *criterio interno* está fundamentado en información obtenida del *algoritmo* de agrupamiento.

Tal como lo señala Milligan (op.cit.; p. 330), han sido sugeridos numerosos índices de naturaleza externa en autores como Rohlf (1974), Mezzich (1975) y Milligan (1978), recomendando este último autor al *estadístico Rand* , índice propuesto por W. M. Rand en 1971. Los autores del presente trabajo consideran que el estadístico Rand reúne características pertinentes que lo hacen atractivo, particularmente, para aquellos investigadores no especializados en estadística matemática y, por ello, se hace la descripción del mismo en la próxima sección de este artículo.

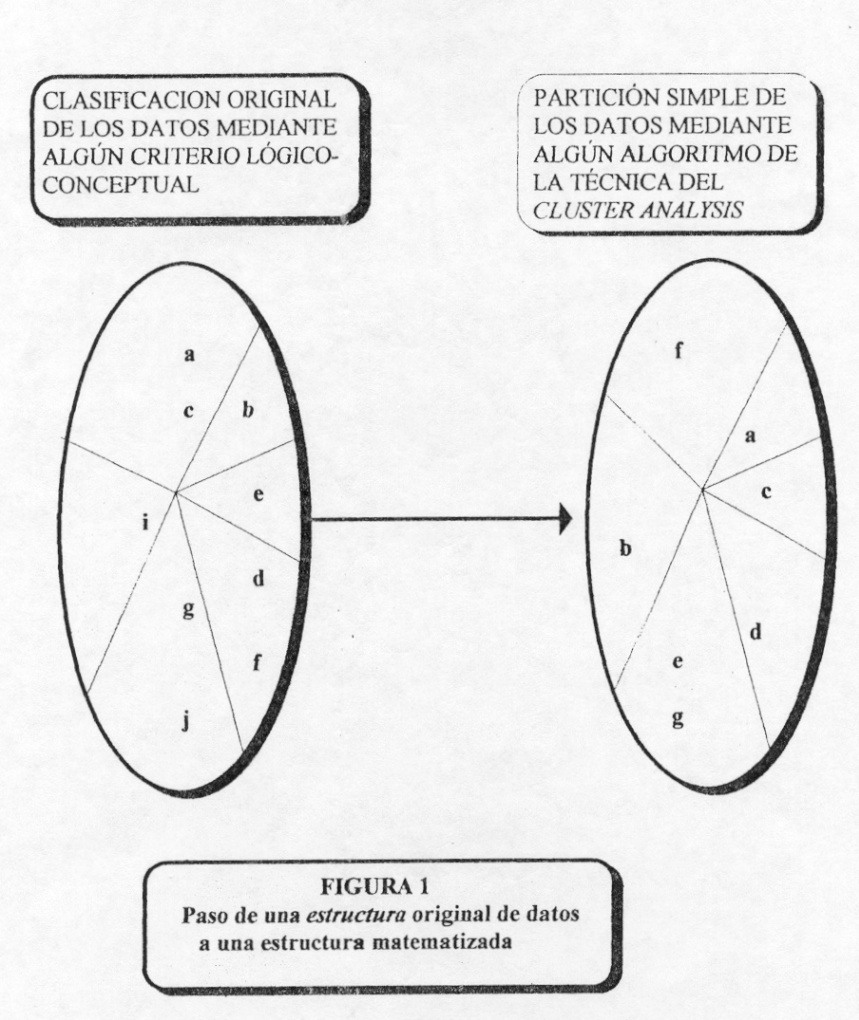


FIGURA 1
Paso de una *estructura* original de datos a una estructura matematizada

EL ÍNDICE RAND

De acuerdo a Milligan (op.cit. p.330), el índice o estadístico Rand (que en adelante se abreviará con el acrónimo I_R) se calcula mediante la siguiente fórmula:

$$I_R = \frac{\sum_{j=1}^{n}\sum_{i>j}^{n}\delta_{i,j}}{\frac{n*(n-1)}{2}} \qquad (1)$$

donde,

$\delta_{i,j}$ … elemento de la i-ésima fila y la j-ésima columna de la matriz Rn

matriz Rn … matriz cuadrada y simétrica, de orden n, que se simbolizará como [Rn]

n …… orden de la matriz cuadrada y simétrica Rn

Para una mejor comprensión de cómo se compila la matriz Rand se ha convenido en presentar las figuras 2 y 3.

El primer paso será construir una *Matriz de Relaciones entre Taxones de Suelo* (matriz izquierda de la figura 2), la cual se estructura de la manera siguiente: tanto en las filas como en las columnas se indica mediante una abreviatura (en este caso de 3 letras) el orden del suelo al cual pertenece cada perfil; por ejemplo, la abreviatura INC en la primera fila significa que el perfil 1 pertenece al orden INCEPTISOL; ULT en la fila 2 implica que el perfil 2 fue clasificado como ULTISOL; de esta manera se completará sucesivamente el indicador del orden hasta completar los 30 perfiles.

Obviamente, según lo escrito precedentemente, la misma secuencia de abreviaturas realizada para las filas se repetirá para los encabezados de las columnas; de este modo, el cruce, por ejemplo, de la fila 1 con la columna 1 establecerá una relación del perfil 1 consigo mismo.

En la parte central o núcleo de la matriz 1 se vaciará la información referente a la conexión que tienen los perfiles entre sí en el sentido de si pertenecen o no al mismo orden, estableciéndose la siguiente regla o criterio: si el perfil de la fila 1 y el perfil de la columna 2 pertenecen al mismo orden se codificará, en el cruce de las líneas que provienen de filas y columnas, el numeral 1. Si ambos perfiles no pertenecen al mismo orden se colocará el símbolo cero (0). Por convención, el cruce de las líneas correspondientes al mismo perfil se codificará también con 0, denotando con ello que no interesa esta solución trivial, es decir, no tiene sentido comparar al perfil consigo mismo; de esta manera, la diagonal principal de la matriz solamente tendrá ceros. Parte de lo descrito para la matriz 1 se recalca en la parte inferior de la misma en la figura 2.

Seguidamente se elaborará la matriz 2 (*Matriz de Relaciones entre Clases de la Partición*), esquematizada en la matriz derecha de la figura 2. Aquí filas y columnas están identificadas por numerales asignados según los resultados del algoritmo matemático de agrupación. El número posible de numerales es siete, desde el numeral 1 hasta el numeral 7, debido a que se requiere llegar con igual número de agrupaciones al que posee el conjunto original (matriz 1); a este respecto cabe señalar que los 30 perfiles de suelos seleccionados constituían un agrupamiento de siete (7) órdenes distintos de suelos.

Al igual que ocurrió con la matriz 1, en la parte central o núcleo de la matriz 2 se compilará la información referente a la relación que tienen las *clases automáticas* entre sí desde el punto de vista si pertenecen o no al mismo subconjunto de la partición simple. Si los perfiles 1 y 2 son de la misma clase, se codificará, en el cruce de las líneas provenientes de filas y columnas, respectivamente, el numeral 1. Si los dos perfiles no son del mismo subconjunto se colocará el símbolo 0 en el cruce de líneas. La diagonal principal de la matriz está constituída por ceros debido a la convención aludida en párrafo anterior.

La matriz Rand se compilará combinando las matrices anteriores según el criterio o la *función* convencional que se describirá a continuación (véase figura 3). Se establecerá la siguiente relación *lógico-matemática*: si cada celda i,j de las matrices 1 y 2 tienen igual numeral, se escribirá el numeral 1 en la correspondiente celda i,j de la matriz Rand. Cualquier otra combinación de numerales para las matrices 1 y 2 implicará llenar con 0 la celda i,j de la matriz Rand.

En resumen, el elemento $\delta_{i,j}$ solamente puede tener las cifras 0 o 1, compilándose el numeral si y sólo si: (i) los *individuos* i y j pertenecen al mismo grupo antes y después de la partición simple; o sea, si dos individuos fueron clasificados *racionalmente* en el mismo taxón y son agrupados matemáticamente, a posteriori, en el mismo subconjunto, entonces se escribirá el numeral 1; (ii) si los individuos i y j no pertenecen al mismo grupo antes y después de la partición matemática, es decir, si dos individuos fueron clasificados racionalmente en grupos diferentes y son agrupados matemáticamente, a posteriori, en clases diferentes entonces se compilará el numeral 1. Naturalmente, se anotará cero, en la matriz Rand, en las situaciones o combinaciones restantes.

Es pertinente aclarar, sobre todo al no versado en simbología matemática, la interpretación de la fórmula del I_R. (a) La doble suma o *sumatoria* ($\Sigma\Sigma$) significa que se sumarán los elementos matriciales tanto por fila como por columna. Por ejemplo, el *operador matemático* Σ con subíndice j, a la izquierda del operador Σ, con subíndice i, manifiesta que primero se totalizará por filas (*horizontalmente*) y luego por columnas (*verticalmente*); (b) dado que en toda matriz simétrica los elementos situados por arriba de la diagonal principal son exactamente iguales a los situados por debajo de dicha diagonal, la relación $i>j$ elimina la posibilidad de que el mismo elemento sea considerado dos veces en el proceso *sumatorio*. Por ejemplo, supóngase una matriz cuadrada y simétrica de orden 5. Operacionalmente se tendría que:

$$\sum_{j=1}^{5}\sum_{i>j}^{5}\delta_{i,j}=\sum_{2}^{5}\delta_{i,1}+\sum_{3}^{5}\delta_{i,2}+\sum_{4}^{5}\delta_{i,3}+\sum_{5}^{5}\delta_{i,4}$$

Si con propósitos de simplificación, simbolizamos la doble suma con la letra S y cada uno de los sumandos arriba señalados como S_i, donde i=1, 2, 3, 4, tendríamos que :

$S= S_1 + S_2 + S_3 + S_4$; el número de sumandos por cada término S_i será de:

En S_1 hay: 4 sumandos
En S_2 hay: 3 sumandos
En S_3 hay: 2 sumandos
En S_4 hay: 1 sumando

→ En total habrá 10 sumandos

Los elementos i,j que se considerarán por cada término sumatorio son:

En S_1…….(2,1) (3,1) (4,1) (5,1)
En S_2…….(3,2) (4,2) (5,2)
En S_3…….(4,3) (5,3)
En S_1…….(5,4)

Entiéndase que, por ejemplo, la simbología (2,1) significa el elemento que está en la celda donde se cruzan la fila 2 y la columna 1. El resto de los números entre paréntesis se interpretarán de manera análoga.

Resulta, de acuerdo a los elementos que se consideran para S_1 que, de la primera columna de la matriz Rand, solamente se toman los elementos desde la segunda fila en adelante. Para la segunda columna, solamente se toman los elementos de la tercera fila en adelante. Para la tercera columna, los elementos de la cuarta fila en adelante y, finalmente, para la quinta columna, sólo el elemento de la cuarta fila. (c) Con respecto al denominador n (n-1)/2 es apropiado hacer el siguiente comentario. Si n es el orden de una matriz cuadrada, el producto n*n nos da el número de elementos de la matriz cuadrada de orden n. Como el número de elementos de la diagonal principal es de n términos , al producto n*n le restaríamos esos n términos, quedando: (n*n)-n= n(n-1). En consecuencia, n(n-1) es el número de elementos que hay en una matriz cuadrada de

orden n pero sin incluir los elementos de la diagonal principal. Como [Rn] es una matriz simétrica, es decir, con los mismos elementos por arriba y por debajo de la diagonal principal, no interesándonos para el cómputo de I_R utilizar valores repetidos, la operación [n(n-1)]/2 significa que sólo se está tomando en consideración el número de elementos de la mitad superior (o inferior) de la matriz Rn. En una matriz cuadrada y simétrica de orden 5, solamente se sumarían 10 elementos porque [5(5-1)]/2=20/2=10. En conclusión, la expresión [n(n-1)]/2 da la cantidad de elementos a considerar en el cómputo del estadístico I_R. (d) ¿Cuál es el propósito del denominador n(n-1)/2 en el cálculo del I_R? Si los n(n-1) elementos de la matriz Rn de orden 5 –la matriz R_5 —fueran iguales a la unidad, se tendría que el numerador sería igual a 10; dado que el denominador también es 10, entonces I_R=1. Esto implica que la fracción que se está analizando da la suma de la parte superior, o inferior, de la matriz Rn cuando todos los elementos de la porción triangular superior o inferior son iguales a 1. En consecuencia, el cociente utilizado para obtener a I_R es la *operacionalización* matemática de la comparación de la suma de valores observados en la matriz Rand con la suma máxima posible en la mencionada matriz; (e) ¿Cómo se interpreta el I_R? Si cada una de las celdas de la porción triangular escogida de la matriz Rand es igual a 1, el numerador de I_R para una matriz , por ejemplo, de orden 5 es igual 10; ello significa que la relación entre individuos según la clasificación racional y según la clasificación automática es la misma. Al aplicar la fórmula del I_R nos da igual a 1; en consecuencia si el I_R=1, ello implica que ha habido una recuperación perfecta de la estructura original de los datos por parte del algoritmo matemático; es decir, I_R=1 significa una tasa de recuperación del 100%. Si ahora la suma del numerador de I_R fuera cero, entonces la clasificación automática tiene una recuperación nula de la estructura original de datos.

En resumen, se puede esquematizar lo anterior así:

Si I_R =

- 0 — necesariamente el numerador del I_R = 0; ello implica total incompatibilidad entre clasificaciones pre y post aplicación del algoritmo matemático
- 1 — necesariamente son iguales numerador y denominador del I_R, por lo que clasificaciones pre y post aplicación del algoritmo son 100% *compatibles o equivalentes*.

En conclusión, el I_R pertenecerá al *intervalo cerrado* [0;1], siendo deseable que el valor muestral esté cercano a 1, lo que indicaría una partición relativamente equivalente a la clasificación racional.

APLICACIÓN DEL ÍNDICE RAND

Con base a una investigación realizada por Fernández y Rodríguez (1995), se obtuvo la información correspondiente a la clasificación taxonómica de 30 perfiles de suelos bajo distintas condiciones naturales; tales perfiles presentaron la siguiente distribución de frecuencias al nivel de orden de suelos:

Número de Grupo	**Número de Casos**	**Orden de Suelos**
1	8	INCEPTISOL
2	12	ULTISOL
3	5	ARIDISOL
4	2	OXISOL
5	1	ENTISOL
6	1	VERTISOL
7	1	ALFISOL

Estos mismos 30 perfiles se agruparon aplicando las distintas estrategias del *Cluster Analysis* (según el *menú* incluido en el *paquete estadístico Statgraphics 2.6 y 4.0*), entre las cuales cabe mencionar: *Single Linkage, Complete Linkage* y *Group Average*. Finalmente, los autores seleccionaron la estrategia *Seeded*, con la que se obtuvo la siguiente partición de siete clases:

Número de Grupo	**Número de Casos**	**Orden según clasificación automática**
1	3	INCEPTISOL
2	8	ULTISOL
3	5	ARIDISOL
4	1	OXISOL
5	11	ENTISOL
6	1	VERTISOL
7	1	ALFISOL

El resultado de la tabla anterior, así como del encabezado de la última columna, tiene la siguiente explicación: la estrategia *seeded* consiste en seleccionar un perfil por cada taxón de la clasificación de suelos para que se agrupen los perfiles restantes (23) y constituyan, finalmente, los 7 grupos correspondientes. En vista de ello, se denominó a cada uno de los grupos de la partición resultante con el nombre del perfil que se usó como *semilla*.

El paso siguiente fue el de comparar cada uno de los perfiles con el resto en dos fases: *antes de la partición simple* y *después de la partición simple*. Con *antes de la partición simple* lo que se determina es cómo se relaciona un perfil dado con el resto de los perfiles según el orden de suelos (véase figura 2 para recordar este procedimiento); con *después de la partición simple* se determina la relación entre perfiles según los resultados de la clasificación automática (véase figura 2).

Si las situaciones antes y después, del perfil seleccionado en relación a otro dado, es la misma se colocará 1 en la celda correspondiente de la matriz Rand (véase figura 3 para comprender el procedimiento). Con esto quiere manifestarse que el numeral 1 implica que se ha conservado la relación taxonómica entre los perfiles seleccionados. Si no sucede así, se compilará en dicha celda el cero. Los resultados generales de la matriz Rand son los que siguen:

Orden de la matriz Rand	30
Número total de elementos	900
Número de elementos en la diagonal principal	30
Número de elementos en la porción inferior (o superior)	435

Al aplicar la *Hoja de Cálculo Lotus 123*, se determinó que:

$$I_R = \frac{335}{435} = 0{,}77$$

CONCLUSIONES

Un I_R de 0,77 significa que se recuperó el 77% de la estructura original; es decir, en el 77% de los casos, cada perfil seleccionado se mantuvo asociado con el comparado tanto en la estructura original como en la partición obtenida con la estrategia *seeded*.

Es menester hacer algunas observaciones finales sobre la metodología y resultados presentados, las cuales se muestran de inmediato.

El I_R, tal como lo aplica Milligan (1980), es con el propósito de evaluar la calidad de distintas estrategias del Análisis de Particiones en cuanto a su capacidad para reproducir o recuperar la estructura original de los datos. Sin embargo, a partir de la experiencia de los autores de este

(1)

MATRIZ DE RELACIONES ENTRE TAXONES DE SUELOS

ORDENES DE SUELOS

		1	2	....	....
		INC	ULT	....	....
1	INC	0			
2	ULT		0		
...	...			0	
...	...				0
29	ARD				
30	ARD				

(2)

MATRIZ DE RELACIONES ENTRE CLASES DE LA PARTICION SIMPLE

CLASES *AUTOMÁTICAS*

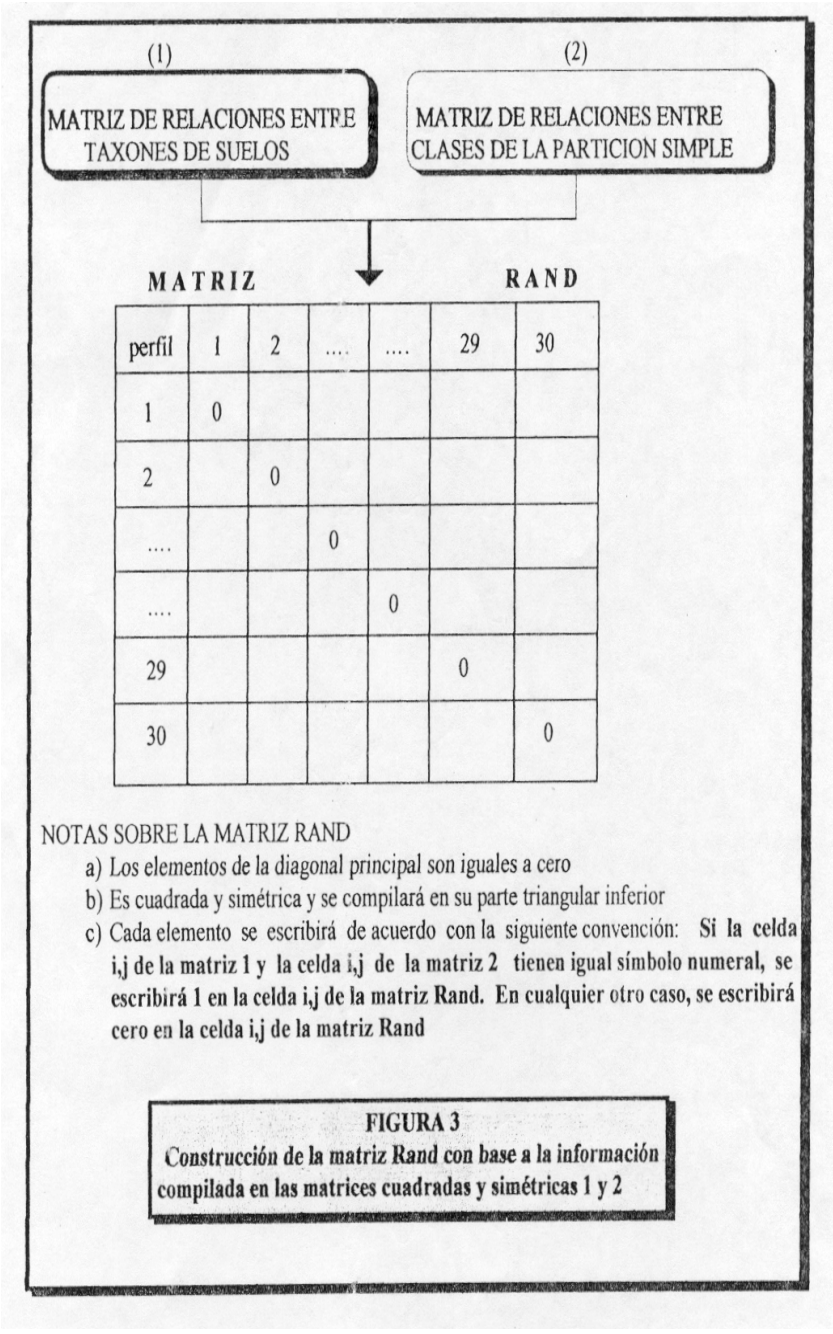
		1	2	...	....	
		1	2			
1	1	0				
2	2		0			
...	....			0		
..	..				0	
29	3					
30	3					

NOTAS:

(1) Todos los elementos de la diagonál principal son iguales a cero
(2) Si dos perfiles pertenecen al mismo Orden, se compilará 1 en la celda donde ellos se *cruzan*
(3) Si dos perfiles no pertenecen al mismo Orden, se compilará cero en la celda donde ellos se cruzan

(1) Todos los elementos de la diagonal principal son iguales a cero
(2) Si dos perfiles pertenecen a la misma clase, se compilará 1 en la celda donde ellos se *cruzan*
(3) Si dos perfiles no pertenecen a la misma clase, se compilará 0 en la celda donde ellos se cruzan

FIGURA 2
Elaboración de las matrices previas a la construcción de la matriz Rand



perfil	1	2	....	....	29	30
1	0					
2		0				
....			0			
....				0		
29					0	
30						0

NOTAS SOBRE LA MATRIZ RAND

a) Los elementos de la diagonal principal son iguales a cero

b) Es cuadrada y simétrica y se compilará en su parte triangular inferior

c) Cada elemento se escribirá de acuerdo con la siguiente convención: **Si la celda i,j de la matriz 1 y la celda i,j de la matriz 2 tienen igual símbolo numeral, se escribirá 1 en la celda i,j de la matriz Rand. En cualquier otro caso, se escribirá cero en la celda i,j de la matriz Rand**

FIGURA 3
Construcción de la matriz Rand con base a la información compilada en las matrices cuadradas y simétricas 1 y 2

trabajo, se avizora que tiene, al menos, dos aplicaciones complementarias, las que calificamos de potencialmente productivas, y en las que actualmente laboramos:

1) El I_R permitiría evaluar si la clasificación racional ha sido elaborada libre de errores.

2) El I_R permitiría evaluar cuán equivalentes son dos sistemas clasificatorios.

En conclusión, con los enfoques aquí esbozados, se plantea que el estadístico o índice Rand, por su sencillez y potenciales aplicaciones, sería un instrumento operacional útil a los geógrafos que trabajan en los problemas de clasificación de las unidades de muestreo, particularmente de aquellas que poseen alguna distribución espacial.

REFERENCIAS BIBLIOGRÁFICAS

Fernández, L. y Rodríguez, J. "Aplicación de la clasificación automática para evaluar la clasificación taxonómica de suelos a nivel de orden". En: *XLV Convención Anual Asovac*. 19-24 de noviembre de 1995. Caracas.

Mezzich, J. E. "An evaluation of quantitative taxonomic methods". (Doctoral dissertation, The Ohio State University, 1975). *Dissertation Abstracts International*, 1975, *36*, 3008-B (University Microfilm No. 75-26, 626).

Milligan, G.W. "An examination of the effect of error perturbation of constructed data on fifteen clustering algorithms". (Doctoral dissertation, The Ohio State University, 1978). *Dissertation Abstracts International*, 1979, *40*, 401B-4011B. (University Microfilms No. 7902188).

Milligan, G.W. "An examination of the effect of six types of error perturbation on fifteen clustering algorithms". *Psychometrika*, 1980, *44*, 325 – 342.

Rand, W. M. "Objective criteria for the evaluation of clustering methods". *Journal of the American Statistical Association*, 1975, *66*, 846 – 850.

Rohlf, F. J. "Methods of comparing classifications". *Annual Review of Ecology and Systematics*, 1974, *5*, 101 –113.