UNIVERSIDAD DE LA REPÚBLICA
Facultad de Ciencias Económicas y de Administración
Instituto de Estadística

Determinación de tipologías de infecciones parasitarias intestinales, en escolares mediante, técnicas de clustering sobre datos binarios

Ramón Alvarez - Fernando Massa
Octubre 2012

Documentos de Trabajo

Serie DT (12 / 05) - ISSN : 1688-6453

Determinación de tipologías de infecciones parasitarias intestinales, en escolares mediante, técnicas de clustering sobre datos binarios

Ramón Alvarez [1]
Instituto de Estadística - Facultad de Ciencias Económicas y de Administración - UdelaR.

Fernando Massa [2]
Instituto de Estadística - Facultad de Ciencias Económicas y de Administración - UdelaR.

RESUMEN

El presente documento describe la metodología y los resultados para la determinación de tipologías de infección parasitaria en escolares mediante técnicas de clustering sobre datos binarios.

En un estudio sobre nutrición y parasitosis intestinal, los diferentes parásitos intestinales evaluados se agruparon en 3 familias: geohelmintos, otros patógenos y no patógenos, lo que genera 3 variables binarias, que serán usadas para la construcción de tipologías de infección. Teniendo en cuenta que se trata de datos binarios los métodos convencionales usados para clustering, donde se maneja una métrica euclideana, no corresponden con lo cual se propone la aplicación de otras técnicas.
Se ensaya un método de cluster probabilístico con una variable latente, estimado mediante el algoritmo *EM*(método 1), un método mixto que combina clustering basado en medidas de entropía con una partición difusa estimada con el algoritmo *c-modes*(método 2), y un tercer método de clustering jerárquico sobre distancias binarias de tipo 'simple matching'(método 3).

Los dos primeros métodos proporcionan clusters difusos y el tercero establece una partición, por lo cual se comparan los resultados a través de un análisis de sensibilidad, donde se da cuenta del comportamiento de los métodos al cambiar la inicialización y el número de iteraciones.
Palabras claves: Algoritmo *EM*, Distancias binarias, Fuzzy clustering.

[1]ramon@iesta.edu.uy
[2]fmassa@iesta.edu.uy

1. Introducción

En el año 2010 la Escuela de Nutrición y Dietética y la Facultad de Medicina, junto al IESTA, se presentaron a un proyecto de Investigación de PIM (Programa de Inserción Metropolitana), para estudiar la asociació entre estado nutricional y la infección por parásitos intestinales de los niños que concurrían a la escuela 307 de la ciudad de Montevideo. A partir del Censo de Talla del año 2002 realizado por la Administración Nacional de Educación Primaria (ANEP) a niños de primer grado de escuelas públicas de todo el país, esta escuela mostró los más elevados porcentajes de retraso de talla: $34,7\,\%$ de retraso moderado y $5,7\,\%$ de retraso grave(ANEP, 2003). De esta manera se planteó un estudio longitudinal descriptivo donde participaron los niños de la escuela 317, a los que se hicieron dos tomas de datos y una intervención entre ambas aplicando medicación para el tratamiento de la parasitosis. Los objetivos seguidos eran

- Conocer el estado nutricional de los escolares
- Conocer la prevalencia de las enteroparasitosis en dichos escolares
- Discriminar la prevalencia de cada uno de los agentes parasitarios hallados
- Relacionar el estado nutricional con la presencia de agentes parasitarios
- Evaluar el estado nutricional luego del tratamiento antiparasitario correspondiente
- Crear una tipología de niños con diferentes perfiles nutricionales y de infección parasitaria.
- Capacitar al personal de la escuela y a las familias de los niños en relación con el control y la profilaxis de las parasitosis intestinales
- Capacitar al personal de la escuela y a las familias de los niños en relación con el consumo adecuado y seguro de alimentos

Para la evaluación de las mediciones antropométricas (peso, y talla relacionados con la edad), se usó la metodología de la OMS de Z scores(de Onis y Blossner, 1997).

Los diferentes parásitos intestinales son evaluados a partir de variables binarias que serán usadas para la construcción de tipologías de infección.
De esta manera al tener variables dicotómicas es necesario considerar metodologías de construcción de grupos que consideren métricas alternativas a la euclideana, con lo cual en la sección 2 se presentan metodologías de cluster probabilístico a partir de variables latente, método mixto que combina clustering basado en medidas de entropía con una partición difusa y método de clustering jerárquico.

2. Metodología

2.1. Clustering basado en un modelo probabilístico - algoritmo *EM*

Siguiendo la metodología propuesta en Moustaki y Papageorgiou (Moustaki y Papageorgiou, 2004), en los modelos de clase latente se asume la existencia de una variable aleatoria Z que consta de G clases. Sea $(x_1, x_2, \ldots, x_p)$ un vector de p variables explicativas con distribución Bernoulli con parámetro ϕ_{jk} dónde el subíndice k hace referencia a cada una de las variables y j a cada uno de los grupos que se introducirán a continuación. Para completar la especificación del modelo es necesario definir las probabilidades τ_j, las mismas suelen llamarse probabilidades a priori. La distribución conjunta de las variables observadas está dada por la siguiente mezcla finita probabilística:

$$f\left(x_i\right) = \prod_{j=1}^{G} \prod_{k=1}^{p} [\tau_j \phi_{jk}^{x_{ik}} (1-\phi_{jk})^{(1-x_{ik})}]^{I(Z_i=j)} \tag{1}$$

Es necesario incluir la variable aleatoria $I(Z_i = j)$ debido a que no se sabe a que grupo pertenece cada observación. De esta manera, es posible plantear el problema de la asignación de individuos a grupos utilizando la distribución de la variable I. Los parámetros del modelo se estiman iterativamente a través del algoritmo EM (Dempster et al., 1977). Este proceso iterativo se repite hasta que el cambio en la log-verosimilitud sea menor que cierta tolerancia pre-especificada. Para este caso, cada iteración de dicho algoritmo, consta de las siguientes etapas:

$$L(\theta|x) = \sum_{i=1}^{n} \sum_{j=1}^{G} I_{(Z_i=j)} \sum_{k=1}^{p} [\log(\tau_j) + x_{ik} \log(\phi_{jk}) + (1-x_{ik}) \log(1-\phi_{jk})] \tag{2}$$

Finalmente la log-verosimilitud de una muestra de n individuos está dada por la siguiente expresión:

1. paso E) $Q(\theta|\theta^t) = E_{Z|X,\theta^t}[L(\theta|x)]$
2. paso M) $\theta^{t+1} = \underset{\theta}{\max}\, Q(\theta|\theta^t)$

Este proceso iterativo se repite hasta que el cambio en la log-verosimilitud sea menor que cierta tolerancia pre-especificada. Cabe señalar que uno de los principales inconvenientes de este algoritmo es su dependencia del valor inicial del vector de parámetros, para hacer frente a este problema en este trabajo se optó por la estrategia de inicializar el algoritmo con distintas valores iniciales y seleccionar los resultados del modelo que obtenga el mayor valor de la log-verosimilitud.
Algunas de las ventajas de este enfoque son las siguientes:

- Dada la naturaleza paramétrica del modelo, se puede seleccionar el número de grupos utilizando criterios de información como el Bayesian Information Criterion (BIC) o el Akaike Information Criterion (AIC).
- A través de la distribución a posteriori de la variable aleatoria Z condicional a los parámetros y a las variables binarias, se puede definir el grado de pertenencia de cada individuo a cada grupo.

$$P(Z_i = j) = \frac{\tau_j \prod_{j=1}^{G} \phi_{jk}^{x_{ik}} (1 - \phi_{jk})^{1 - x_{ik}}}{\sum_{j=1}^{G} \tau_j \prod_{j=1}^{G} \phi_{jk}^{x_{ik}} (1 - \phi_{jk})^{1 - x_{ik}}} \tag{3}$$

- Mediante los parámetros del modelo se pueden calcular frecuencias esperadas que, comparadas con las frecuencias observadas, indiquen el grado de ajuste del modelo.

2.2. Clustering fuzzy - algoritmo *Fuzzy C-modes*

Acorde a la metodología propuesta por (Tsekouras et al., 2005) en un artículo donde presenta como clasificar atributos categóricos en la cultura para inmigrantes, el algoritmo *Fuzzy C-modes* provee una representación flexible de la estructura de los datos ya que cada individuo pertenece a más de un cluster con diferentes grados de participación. A diferencia del modelo de cluster probabilístico, este método requiere el uso de una distancia. En el artículo antes citado se utiliza la distancia denominada *simple matching*, que se consiste en

$$D(x_i, x_l) = \sum_{k=1}^{p} \delta(x_{ik}, x_{lk}) \tag{4}$$

donde

$$\delta(x_{ij}, x_{lj}) = \begin{cases} 0 & \text{si } \ x_{ik} = x_{lk} \\ 1, \text{en otro caso} \end{cases} \tag{5}$$

Luego, el algoritmo se basa en minimizar la siguiente función objetivo:

$$J_m(U, V) = \sum_{i=1}^{n} \sum_{k=1}^{G} (u_{ji})^m D(x_i, v_k) \tag{6}$$

donde n es el número de individuos, G el número de grupos, m es el parámetro de incertidumbre (fuzziness parameter), v_k es el centro del cluster c y u_{ik} es el grado de pertenencia del individuo i al grupo k.

Para poder encontrar el número adecuado de clusters, (Tsekouras et al., 2005) proponen un método en 3 etapas tal como se muestra en la figura que sigue
Originalmente (Tsekouras et al., 2005) plantean un algoritmo que aparece en la figura 1, donde construyen los clusters mediante medidas de entropía como se detalla más adelante para el paso 1.
El valor de *entropía* entre 2 objetos de tipo categórico, como se definieron en 4 se puede establecer como

$$H_{il} = -E_{il} log_2(E_{il}) - (1 - E_{il}) log_2(1 - E_{il}) \tag{7}$$

donde E_{il} es la *similaridad* entre 2 objetos que se define como

$$E_{il} = exp\{-aD(x_i, x_l)\} \tag{8}$$

con $a \in (0, 1)$. De esta manera la *entropía* total de x_i con respecto al resto de los individuos es

$$H_i = \sum_{l=1, l \neq i}^{n} [-E_{il} log_2(E_{il}) - (1 - E_{il}) log_2(1 - E_{il})] \tag{9}$$

(Tsekouras et al., 2005) establecen que:

> La entropía total de un objeto será pequeña cuando varios objetos vecinos lo rodean, con lo cual un objeto con entropía pequeña es un buen candidato a centro del cluster.

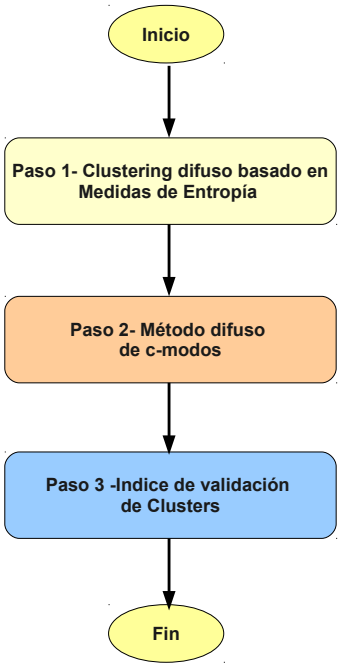

Figura 1: Algoritmo de conformación de los grupos para datos categóricos

Para el caso de este trabajo sobre parasitosis en la escuela de Montevideo, los autores deciden hacer una modificación del algoritmo considerando solamente los primeros 2 pasos del mismo, donde a su vez se plantea una variante para el paso 1. Para la elección de los centros de los clusters, proponen una partición con *grados de pertenencia* u_{ik} en forma inicial que se genera simulando a través de una distribución *dirichlet* $p(u_{ik}) = Dirichlet(u_{ik}|\tau_1, ..., \tau_G)$.

2.3. Clustering con métodos jerárquicos

Los métodos jerárquicos se caracterizan por generar una serie de particiones encajadas y, a diferencia de los dos métodos anteriores requieren la definición de una distancia. Dado el carácter binario de las variables a utilizar, la distancia seleccionada debe de ser del tipo que se presenta a continuación.
Supongamos que se tiene 2 atributos binarios, para los cuales se arma la siguiente tabla tetracórica

a	b	**a+b**
c	d	**c+d**
a+c	**b+d**	**a+b+c+d**

- a = número de individuos que comparten ambos atributos
- b = número de individuos que no tienen el primer atributos pero si el segundo
- c = número de individuos que tienen el primer atributos pero no el segundo
- d = número de individuos que no tienen ningunos de los atributos
- N = a+b+c+d

A partir de esta tabla se pueden armar muchas distancias de tipo simétricas y asimétricas, de las cuales para este trabajo los autores consideran la de simplematching (Jurasinski y with contributions from Vroni Retzer, 2012)

$$sim = \frac{a+d}{a+b+c+d} \tag{10}$$

Esta es una distancia ampliamente usada en ecología, aunque ya existen aplicaciones en el campo de la salud y de la economía (Alvarez et al., 2011). Pese a ser sencilla de

interpretar, esta distancia posee la desventaja de que dos pares de individuos con distintas configuraciones de ceros y unos pueden estar a la misma distancia. Sobre esta matriz de distancia es que luego se aplica el algoritmo de *Ward*(Kaufman y Rousseeuw, 1990) de carácter agregativo que busca optimizar, en cada etapa, la dispersión de las clases de la partición obtenida por agregación de individuos o grupos tal como se presenta en la sección 2.3.1.

2.3.1. Método de Ward

Al descomponer la variación total en variación en los grupos (*within*) y variación entre los grupos (*between*) y estar frente a una partición dada, el método unirá aquellos grupos que produzcan el efecto de hacer mínima la variación *within* en la nueva partición.

$$T = W + B \tag{11}$$

Donde T es la matriz de varianzas y covarianzas del total, W la matriz de varianzas y covarianzas dentro de los grupos y B la matriz de varianzas y covarianzas entre grupos. En este caso para determinar con que cantidad de grupos trabajar, existen varias reglas de detención, de las cuales se presentan algunas

R cuadrado: Establece la relación entre la variación explicada y la variación total, donde la variación explicada la representa la estructura de grupos hallada en cada nivel.

$$R^2 = 1 - \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{j=1}^{J}(x_{(ij(k))} - \overline{x}_{kj})^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}(x_{(ij)} - \overline{x}_{j})^2} \tag{12}$$

En cada etapa de particiones encajadas se observa el valor del indicador y el incremento que se produce en el mismo al pasar de k grupos a $k+1$ grupos. Si el aporte deja de ser significativo, se opta por trabajar con k grupos.

Regla de Calinski (llamada Pseudo F):

$$pseudoF = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \tag{13}$$

donde el estadístico $pseudoF$ no tiene distribución F pero del cual empíricamente se han determinado algunas reglas que contribuyen a su utilización:

- si el indicador crece monótonamente al crecer el número de grupos $k \Rightarrow$ no se puede determinar una estructura clara;

- si el indicador disminuye monótonamente al crecer el número de grupos $k \Rightarrow$ no se puede determinar claramente la estructura de grupos, pero se puede decir que existe una estructura jerárquica;
- si el indicador crece, llega a un máximo y luego decrece $\Rightarrow$ la población presenta un número definido de grupos en ese máximo.

Test de Duda-Hart y pseudo t^2: El *pseudo* t^2 al igual que el *pseudo* F es un indicador útil para determinar el número de grupos, pero no tiene distribución exacta *t-student*. Está relacionado con el indicador planteado por Duda-Hart, el que compara las trazas de las matrices de varianzas intragrupales G y L con la traza de la matriz de varianzas que surge al unir los grupos G y L.

$$Duda - Hart = DH = \frac{trW_G + trW_L}{trW_{GL}} \tag{14}$$

Lo que se intenta con estos indicadores es determinar la importancia de fusionar dos grupos. En cada paso considera los candidatos a unirse. Se trata de determinar en cada paso si la disminución en la suma de cuadrados residuales (variación intragrupos, o variación en los grupos) como resultado de pasar de k a $k+1$ grupos es significativa o no. Esto significa que el incremento en la heterogeneidad al unir los grupos es muy grande y por tanto no es conveniente su unión.

3. Aplicación

Tal como se presentó en la sección 1 la tabla de datos con la que se trabajó está formada por los 79 individuos para los que se tenía resultados para todos los exámenes coproparasitarios. De esa manera partiendo de 11 variables de infección por parásitos se agregan en tres variables binarias que indican la presencia/ausencia para 3 familias de parásitos intestinales; geohelmintos, otros patógenos y otros no patógenos; a estas 3 nuevas variables se aplicaron los tres métodos de clustering.

4. Resultados

Los resultados se obtuvieron aplicando funciones desarrolladas en R (R Development Core Team, 2012), por los autores de este trabajo para el caso del método 1 de la sección (2.1) y método 2 de la sección (2.2), mientras que para el método 3 se utilizó la librería *cluster* (Maechler et al., 2012) y la librería *simba* (Jurasinski y with contributions from

Vroni Retzer, 2012). Para el método 1 *E-M* se estima el modelo que genera los grupos con una tolerancia $tol = 1e-3$ y número máximo de 500 iteraciones lo que lleva 30 segundos en promedio de tiempo de máquina. Se muestra en la figura 2 como queda la situación para una corrida en particular donde el algoritmo converge en 114 iteraciones.

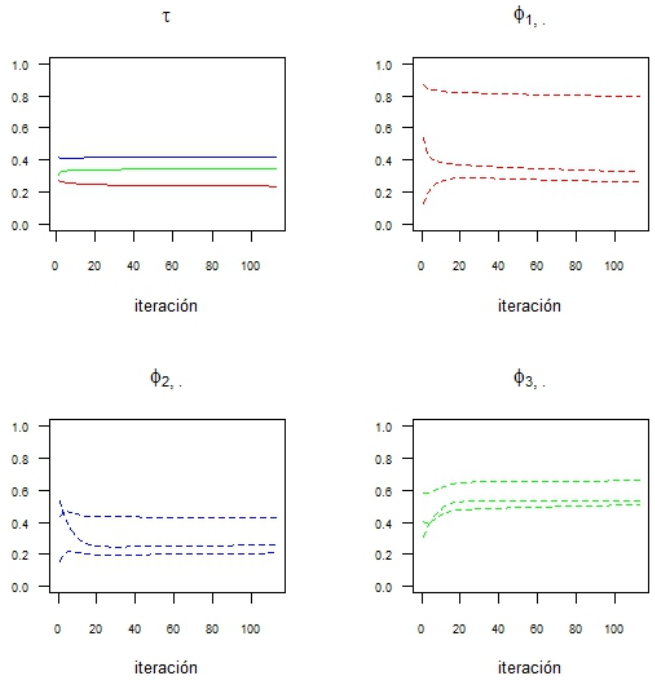

Figura 2: Estimación del número de grupos mediante método 1 (EM)

El algoritmo de estimación del modelo se ejecuta 100 veces para evaluar la estabilidad del mismo, con lo cual finalmente el número de grupos se determina utilizando el criterio BIC en la verosimilitud del modelo estimado por el algoritmo *E-M* como se ve en la figura 7, donde lo indicado es $G = 3$, donde está el mínimo. Se decidió aplicar los otros dos algoritmos con la misma cantidad de grupos de modo que los resultados de las tres técnicas fuesen comparables.
Para el método 2 (c-modes) se presenta en el cuadro 1 las diferentes simulaciones hechas, donde se puede ver como a medida que se incrementa el número de inicializaciones el mínimo de la función objetivo de la ecuación 6 desciende en forma monótona, aunque la variabilidad no decrece.
A su vez para evaluar la calidad de los grupos formados mediante el algoritmo *Fuzzy c-modes* se estimaron 2 índices que forman parte del paso 3 del algoritmo de la figura 1 y que sirven para saber

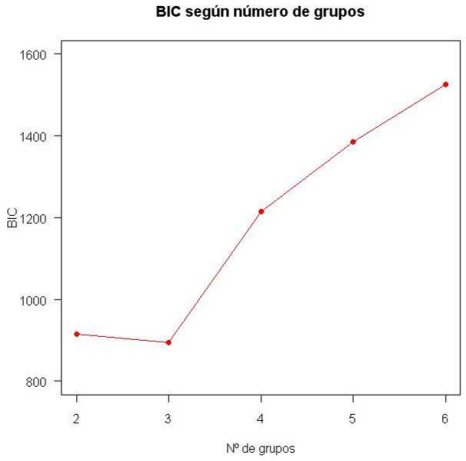

Figura 3: Conformación de los grupos según los tres algoritmos

Inicializaciones	10	20	30	40	50	60	70	80	90	100
promedio	23.1	20.8	19.7	19.27	19.2	19.00	19.1	19.1	19.00	18.9
D.E.	2.5	2.5	1.30	0.5	0.52	0.61	0.50	0.57	0.61	0.64
N95SUP	28.1	25.8	22.2	20.2	20.2	20.2	20.1	20.3	20.2	20.2
N95INF	18.1	15.8	17.1	18.2	18.2	17.8	18.1	18.00	17.8	17.7

Cuadro 1: Variación de la función a minimizar en el algoritmo C-modes

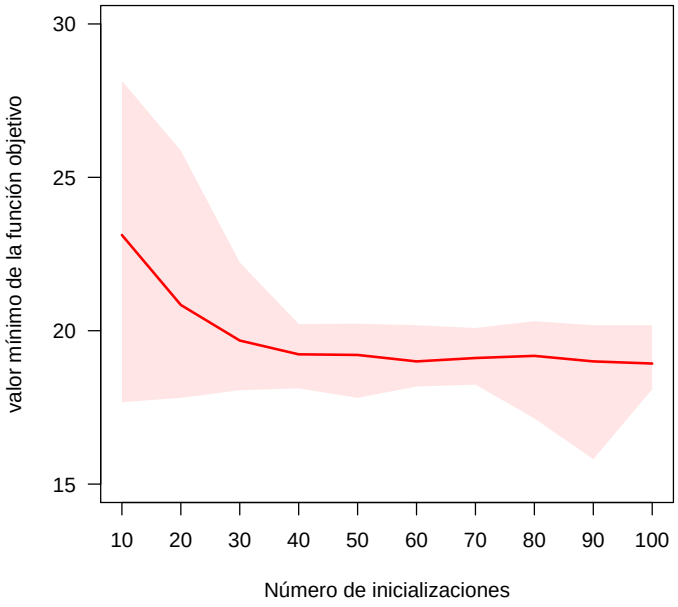

Figura 4: Variación de la función a minimizar en el algoritmo *Fuzzy C-modes*

- Índice de compacidad (IC), que mide que tan compactos son los grupos difusos y que se determina de la siguiente manera

$$(IC) = \sum_{i=1}^{n} \frac{\sum_{k=1}^{G} (u_{ik})^m D(x_i, v_k)}{n_i} \tag{15}$$

- Índice de separación difusa (ISD) que permite saber el grado de separación que se logra entre centros del conjunto difuso

$$(ISD) = \sum_{k=1}^{G} \sum_{l=1, l\neq k}^{m} (u_{ik})^m D(v_k, v_l), \ (l \neq i) \tag{16}$$

De esta manera (Tsekouras et al., 2005) definen para el paso 3 del algoritmo un **índice de validez** (IV) que lo definen como

$$(IV) = \frac{(IC)}{(ISD)} = \frac{\sum_{i=1}^{n} \frac{\sum_{k=1}^{G} (u_{ik})^m D(x_i, v_k)}{n_i}}{\sum_{k=1}^{G} \sum_{l=1, l\neq k}^{m} (u_{ik})^m D(v_k, v_l)} \tag{17}$$

Para la aplicación sobre los datos presentados en 3 se presenta el índice de compacidad (IC) para los grupos, del cual se ve su comportamiento en función del número de grupos considerado y la cantidad de inicializaciones que se genera en cada corrida del algoritmo. Los otros 2 índices de separación (ISD) y de validez (IV) no se reportan ya que para la mismas simulaciones hechas para (IC) variando el número de grupos de $(2-5)$, aparecen valores indeterminados.

Por último se construye la tipología de grupos mediante el método 3 jerárquico usando la distancia de 'simple-matching', donde se considera 3 grupos para comparar con el método 1 y método 2; el dendrograma muestra la distancia donde se corta el árbol de aglomeración

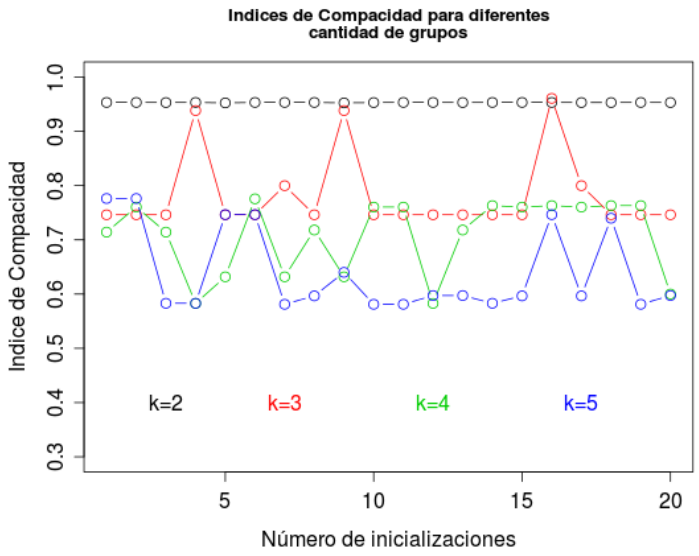

Figura 5: Indices de Compacidad para diferentes cantidad de grupos para algoritmo *Fuzzy C-modes*

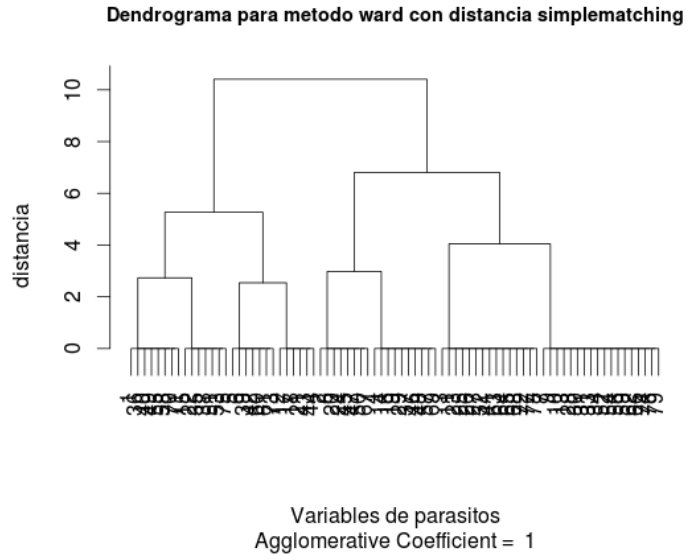

Figura 6: Dendrograma para método Ward con distancia simple-matching

Para evaluar y comparar las 3 tipologías construídas mediante los diferentes métodos se crea una variable que combina las variables binarias de presencia de parásitos, de la siguiente manera:

- un 1 en la primera posición indica presencia de *geohelmintos*
- un 1 en la segunda posición indica presencia de *otros patógenos*
- un 1 en la tercera posición indica presencia de *no patógenos*

Combinación	geo	op	nop	Total
1	0	0	0	15
2	1	0	0	18
3	0	1	0	8
4	1	1	0	10
5	0	0	1	7
6	1	0	1	8
7	0	1	1	6
8	1	1	1	7

Cuadro 2: Clasificación de los individuos de acuerdo a las variables de parásitos

Teniendo en cuenta esa variable que muestra como se combinan las variables de parásitos se cruza con las 3 tipologías obtenidas con los métodos presentados, donde las filas en color muestran los grupos 'puros', es decir de niños que solamente tienen un tipo de parásito.
En el cuadro 4 se puede observar el tamaño relativo de cada grupo y la media condicional de cada variable.
Puede observarse como las metodologías 'difusas' generan particiones similares, en las cuales se destaca un grupo con alta prevalencia de geohelmintos (grupo 2). Por otro lado, los grupos creados mediante al algoritmo de *Ward* ponen de manifiesto grupos mas "puros", como ejemplo el grupo 1 donde se aislaron los niños con parásitos exclusivamente *no patógenos*.Los resultados del cuadro anterior se pueden ver en la figura 7.

En cuanto a la bondad de ajuste del método $E-M$, fue necesario calcular las frecuencias esperadas bajo el modelo. Esta información se presenta en el cuadro 5.
La información presentada en paréntesis corresponde a las frecuencias esperadas, mientras que la dispuesta sin paréntesis representa las frecuencias efectivamente observadas.
El estadístico de Pearson asociado a esta tabla es de $6{,}2\text{x}10^{-4}$ con un *p-valor* asociado de

Variables	EM.1	EM.2	EM.3	FUZ.1	FUZ.2	FUZ.3	W.1	W.2	W.3
000	0	0	15	0	0	15	0	0	15
001	7	0	0	0	0	7	7	0	0
010	0	0	8	0	0	8	0	8	0
011	6	0	0	6	0	0	6	0	0
100	0	18	0	0	18	0	0	0	18
101	8	0	0	5	3	0	8	0	0
110	0	10	0	3	7	0	0	10	0
111	7	0	0	7	0	0	7	0	0
Total	28	28	23	19	30	30	28	18	33

Cuadro 3: Composición de las 3 agrupaciones según las 3 variables binarias combinadas

		tamaño	geohelmintos	otros patógenos	no patógenos
E-M	grupo 1	0.36	0.55	0.49	0.81
	grupo 2	0.26	0.96	0.34	0.07
	grupo 3	0.38	0.38	0.33	0.12
C-Modes	grupo 1	0.39	0.67	0.77	0.72
	grupo 2	0.23	0.84	0.23	0.20
	grupo 3	0.38	0.13	0.27	0.24
Ward	grupo 1	0.35	0.53	0.46	1.00
	grupo 2	0.23	0.55	1.00	0.00
	grupo 3	0.42	0.54	0.00	0.00

Cuadro 4: Medias de cada variable binaria condicional a cada grupo

		No patógenos			
Geohelmintos	**Otros patógenos**	no		si	
no	no	15	(14,99)	7	(7.03)
	si	8	(8,00)	6	(5,96)
si	no	18	(18,00)	8	(7,97)
	si	10	(9,99)	7	(7.03)

Cuadro 5: Frecuencias Observadas y Esperadas en los Grupos Parasitarios por método *EM*

0,98, lo que indica que este modelo describe adecuadamente los datos.

Por último a modo de resumen se presenta la figura 7, donde a través de los gráficos de radar se puede ver la composición de los diferentes grupos.

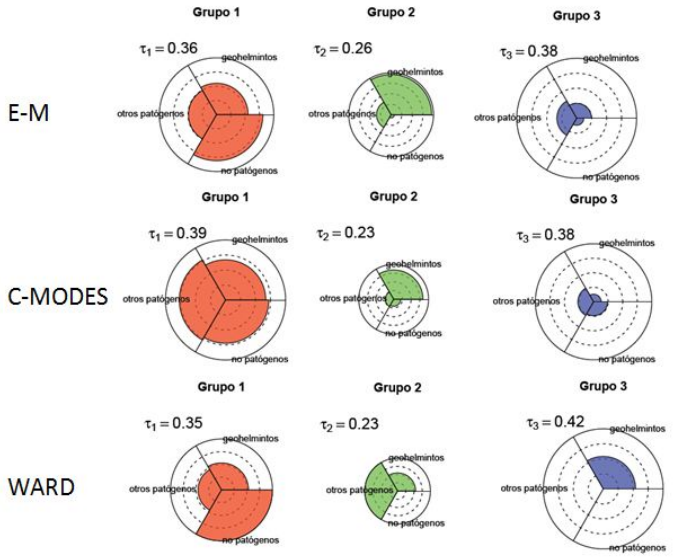

Figura 7: Conformación de los grupos según los tres algoritmos

5. Conclusiones y Futuros pasos

Como conclusiones hasta el momento se puede decir que los 3 métodos considerados e implementados con este ejemplo, son relativamente poco costosos en tiempo de procesamiento y consumo de memoria. Sin embargo esto es en virtud de que el volumen de datos no es excesivo, ya que son 3 variables y 79 observaciones, con lo cual los autores de este trabajo van a probar con otros datos mas voluminosos en variables y en cantidad de observaciones.

Si se tienen n y j variables binarias, en principio la cantidad de grupos que se pueden hacer como máximo son 2^j que puede ser muy diferente de n; en este caso se tenían $2^3 = 8$, que son las que aparecen en los cuadros 3. Por lo tanto el problema puede cambiar si las 3 variables binarias que se crearon agregándolas de acuerdo al tipo de infección que producen, se conservan como tales, que para este caso eran 11 en total .

Considerando entonces los métodos propuestos, se puede decir que los 3 métodos brindan diferentes soluciones, que son flexibles en el sentido que permiten cambiar el número de grupos creados en base a reglas de detención como para el caso de los métodos jerárquicos (método 3), a través de bondad de ajuste para el $E - M$ (m'etodo 1) y por otro lado uno de los métodos posibilita obtener una partición difusa con grados de pertenencia, lo que es una ventaja ya que muchas veces las técnicas de clustering convencionales generan grupos donde reunen observaciones que podrían pertenecer a otro grupo.
En cuanto a la decisión de con cuantos grupos difusos quedarse, los autores proponen seguir investigando porque se producen problemas con los índices de separación (ISD) y de validez (IV), ya que en las simulaciones que se hicieron cambiando el número de grupos para $G = 4$ y $G = 5$ grupos, esos índices en algunas de las 20 inicializaciones no está definido. A su vez para el caso del (IC) este muestra un comportamiento que si bien fluctúa a lo largo de las diferentes inicializaciones del algoritmo, es en general menor cuando crece el número de grupos, lo que es un inconveniente, ya que es deseable que la compacidad sea mínima para el G propuesto.

Por último se propone evaluar diferentes estadísticos que permitan decidir el número de grupos adecuado por método y que luego se puedan comparar resultados, y no como se hizo para este trabajo donde se fijó en $G = 3$

Referencias

Alvarez, R., Canavesi, J., Castrillejo, A., y Massa, F. (2011). Reconstrucción del inse en una encuesta sanitaria poblacional.

ANEP (2003). Tercer censo nacional de talla en niños de primer grado escolar. Technical report, Administración Nacional de Educación Pública. Consejo de Educación Primaria.

de Onis, M. y Blossner, M. (1997). Global database on child growth and malnutrition. Technical report, World Health Organization.

Dempster, A. P., Laird, N. M., y Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39:1–38.

Jurasinski, G. and with contributions from Vroni Retzer (2012). *simba: A Collection of functions for similarity analysis of vegetation data*. R package version 0.3-4.

Kaufman, L. y Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., y Hornik, K. (2012). *cluster: Cluster Analysis Basics and Extensions*. R package version 1.14.2 — For new features, see the 'Changelog' file (in the package source).

Moustaki, I. y Papageorgiou, I. (2004). Latent class models for mixed variables with applications in archaeometry. *Elsevier Computational Statistics & Data Analysis*, page 17.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Tsekouras, G., Papageorgiou, D., Kotsiantis, S., Kalloniatis, C., y Pintelas, P. (2005). Fuzzy clustering of categorical attributes and its use in analyzing cultural data. *World Academy of Science, Engineering and Technology*, 1:87–91.