APLICACIÓN DE ANÁLISIS DE REDES PARA LA ELABORACIÓN DE PERFILES EPIDEMIOLÓGICOS EN ESTUDIOS SANITARIOS

RAMÓN ÁLVAREZ-VAZ [a], FERNANDO MASSA [a], SUSANA LORENZO-ERRO [a,b]

[a] Instituto de Estadística
Universidad de la República de Uruguay
e-mail: ramon@iesta.edu.uy, fmassa@iesta.edu.uy

[b] Servicio de Epidemiología y Estadística, Facultad de Odontología
Universidad de la República de Uruguay
e-mail: susana.of.lorenzo@gmail.com

Habitualmente en los estudios epidemiólogicos se suele trabajar con variables binarias que muestran la presencia de determinadas enfermedades, las que a su vez se asocian con otro conjunto otras enfermedades, denominadas comorbilidades, y que también se miden a través de variables binarias y que en general se asumen como factores de riesgo de las primeras. En el ámbito de estos estudios existen situaciones donde se manejan enfermedades no transmisibles (ENT), en particular en salud bucal, donde ambos tipos de variables pueden ser intercambiables en cuanto a quien hace el rol de factor de riesgo.
Teniendo en cuenta esta situación se propone usar el análisis de redes para la determinación de tipologías de encuestados en base a los atributos binarios, de forma de obtener perfiles epidemiológicos bien diferenciados.
Los datos utilizados corresponden a un estudio en personas que demandan atención en la Facultad de Odontología-UDELAR durante el período 2015-2016. A través del análisis de redes (AR), a partir de las variables se construye la matriz de adyacencias sobre la que se aplican una batería de métricas (*closeness*, *betweenness*, *modularity*, *clustering*) sobre los nodos y enlaces, que permite detectar comunidades.
Las comunidades creadas mediante el AR a través del uso de diferentes algoritmos de búsqueda, como el de *fast greedy* o de *random walk* o de particionado espectral, se usan para evaluar el cambio en las proporciones de las variables analizadas (patologías o factores de riesgo) cuando se consideran en forma global y al interior de cada comunidad.

Keywords: Análisis de redes, clustering, factores de riesgo, variables binarias.

1. Introducción

Las enfermedades no transmisibles (ENT), en las que pueden agruparse enfermedades como las cardiovasculares, diabetes, cáncer y enfermedades respiratorias crónicas, son actualmente la causa de mortalidad a nivel mundial más importante con un peso de $63\,\%$ de las muertes globales, y con una característica extra y es que casi $40\,\%$ de estas muertes se producen entre los 30 y 70 años, con la consecuente carga social al ser el tramo de edad mas relevante donde están las personas económicamente activas, [21]. A su vez la carga de este tipo de enfermedades en los países con bajos y medianos ingresos es del $86\,\%$ de las muertes prematuras [21]. Este conjunto de enfermedades es a su vez responsable de un gran aumento de la discapacidad en varios países del mundo, donde en particular para los de menores ingresos se producen a edades más tempranas, lo que se traduce en discapacidades por períodos más prolongados previo a que sobrevenga la muerte.

En este conjunto de enfermedades es fundamental el papel que juegan los estilos de vida que se relacionan con aspectos como alimentación inadecuada, [24], [8], [10], [1], el sedentarismo, consumo nocivo de alcohol, y el consumo de tabaco. Estos factores a su vez actúan en forma directa o indirecta, creando otros factores de riesgo como son la obesidad, los trastornos del metabolismo de los hidratos de carbono, la hipertensión arterial (HTA) o las dislipemias, [21]. Una preocupación a nivel de la salud pública mundial es tratar de modificar los estilos de vida pasibles, mediante programas preventivos de manera

de lograr una disminución importante en el número de muertes prematuras. Este tipo de problema de la salud pública tiene un impacto muy grande en el desarrollo macroeconómico de los países tal como consigna la OMS y el Foro Económico Mundial y donde se sostiene que en un escenario en que se mantengan estáticos los niveles de intervención, y las cifras de ENT continúen su ritmo de crecimiento, la pérdida económica acumulativa a causa de estas patologías en los países con ingresos medios y bajos superarán los $U\$S$ 7 trillones en el período 2011-2025.
En el contexto de los estudios epidemiológicos donde se indaga por las ENT es práctica habitual trabajar con variables binarias que reflejan la presencia de determinadas enfermedades, las que a su vez se asocian con otro conjunto de enfermedades, denominadas comorbilidades, medidas también a través de variables binarias y que en general se asumen como factores de riesgo de las primeras. En el ámbito de los estudios epidemiológicos existen situaciones donde se manejan ENT, en particular en salud bucal, donde ambos tipos de variables pueden ser intercambiables en cuanto a quien hace el rol de factor de riesgo.

En este trabajo el objetivo es obtener perfiles epidemiológicos bien diferenciados en base a los atributos binarios partiendo de un conjunto de variables, sin discriminar cuales son variables de respuesta, proponiendo la creación de grupos mediante la siguiente estrategia:

1. a través del análisis de redes (AR), a partir de las variables se construye la matriz de adyacencias

2. sobre esta matriz se aplican una batería de métricas (*closeness*, *betweenness*, *modularity*, *clustering*) sobre los nodos y enlaces, que permite detectar comunidades.

El trabajo está organizado de la siguiente forma: en la sección 2 se presentan las técnicas a aplicar y en la sección 3 se presentan brevemente en que consiste el problema en estudio y los datos que se utilizan, para luego mostrar los resultados en la sección 4, discutir los hallazgos en 5, para terminar en la última sección 6, donde se presentan las conclusiones y futuros pasos.

2. Metodología de Análisis de Redes

En esta sección se presentan muy brevemente las diferentes métricas que se usan para la caracterización de las redes sociales, habitualmente usado en AR. Para la presentación de las mismas se seguirá la notación de del libro 'Statistical Analysis of Network Data with R' [13], [15] aunque textos seminales como [22], [4] son una guía también a seguir.

Antes de presentar algunas de las métricas más relevantes para describir una red, es necesario definir conceptos básicos. Una red o grafo es una estructura matemática, la que está formada por 2 tipos de elementos: *nodos* y *enlaces*, donde los nodos pueden ser personas, variables o alguna otra entidad y los enlaces son las relaciones que existen entre los nodos. Se escribe como $G(V, E)$, donde V son los nodos y E los enlaces.

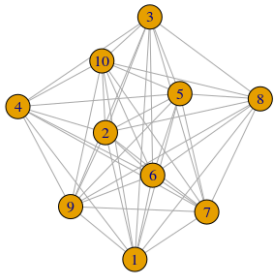

Fig. 1. Ejemplo de red para 10 personas

Si se observa la Figura 2 se ve que el nodo 8 está conectado con los nodos $(3, 10, 5, 2, 9, 1, 6, 7)$, mientras que el nodo 4 se conecta con todos los nodos salvo el 8, como aparece más abajo.

```
1--2  1--3  1--4  1--5  1--6  1--7  1--8  1--9
1--10 2--3  2--4  2--5  2--6  2--7  2--8  2--9
2--10 3--4  3--5  3--6  3--7  3--8  3--9  3--10
4--5  4--6  4--7  4--9  4--10 5--6  5--7  5--8
5--9  5--10 6--7  6--8  6--9  6--10 7--8  7--9
7--10 8--9  8--10 9--10
```

En el contexto de este trabajo, que se presenta en la sección 3, los nodos son personas mientras que los enlaces surgen de considerar si esas personas comparten o no ciertas variables, que en este caso son patologías y hábitos de vida.

Para entender la descripción que se hace del problema desde la perspectiva del AR, es primordial presentar un conjunto de métricas que sirve resumir la información, caracterizar la estructura de la red, a través de lo que se conoce como *topología* del grafo o de la red.

2.1. Grados de los vértices. El grado d_v de un vértice v de un grafo $G(V, E)$ es el número de aristas en E incidentes sobre V. A partir de esta medida se puede definir f_d como la fracción de vértices de $v \in V$ con grado $d_v = d$. El conjunto $\{f_d\}d \geq 0$ es lo que se llama *distribución de grados* de G.

Para las redes ponderadas, una generalización útil del grado es la noción de *Fuerza de vértice* que se obtiene simplemente sumando los pesos de los bordes de un vértice dado.

2.2. Centralidad de los vértices. Las medidas de centralidad de intermediación tienen por objeto resumir en qué medida un vértice se encuentra 'entre' otros pares de vértices [11] **(Betweenness centrality)**

$$c_B(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \quad (1)$$

Donde $\sigma(s,t|v)$ es el número total de caminos más cortos entre s y t que pasan a través de v, y $\sigma(s,t)$ es el número total de caminos más cortos entre s y t (independientemente de si pasan o no por v). Esta medida de centralidad puede rescalarse al intervalo $[0,1]$ mediante un factor de $(N_v - 1)(N_v - 2)$ / 2, siendo N_v el número de vértices del grafo $G(V,E)$.

Las medidas de centralidad o proximidad intentan capturar la noción de que un vértice es 'central' cuenado este se encuentra de 'cerca' de muchos vértices [11], [5]. El enfoque estándar, introducido por [20], es definir a la centralidad como una medida que varía inversamente a la distancia total de un vértice de todos los demás **(Closeness centrality)**

$$c_{CL}(v) = \frac{1}{\sum_{u \in V} dist(v,u)} \quad (2)$$

donde $dist(v,u)$ es la distancia geodésica entre los vértices $u, v \in V$. Para comparar entre otras medidas de centralidad, esta medida se puede rescalar al intervalo $[0,1]$, a través de la multiplicación por un factor $N_v - 1$.

Finalmente, otras medidas de centralidad se basan en nociones de 'prestigio' o 'rango'. Es decir, buscan capturar la idea de que cuanto más centrales sean los vecinos de un vértice, más central es el vértice en sí mismo. Estas medidas pueden expresarse en términos de vectores propios de soluciones de sistemas lineales de ecuaciones.

De acuerdo a [2], [3]

$$C_{E_i}(v) = \alpha \sum_{\{u,v\} \in E} C_{E_i}(u) \quad (3)$$

El vector $C_{E_i} = (C_{E_i}(1), ..., C_{E_i}(N_v))^T$ es la solución al autovalor para $AC_{E_i} = \lambda^{-1} C_{E_i}$, donde A es la matriz de adyacencia del grafo $G(V,E)$. Bonacich sostiene que una elección óptima de α^{-1} es el mayor autovalor de A, y por lo tanto C_{E_i} es el autovector correspondiente. Cuando G es un grafo no dirigido, el valor propio más alto de A será simple y su autovector tendrá valores distintos de cero y del mismo signo.

2.3. Descripción de los enlaces. Se puede extender la idea de intermediación para los enlaces, aspecto que se denomina (Edge betweenness centrality) y que es una extensión de la intermediación de nodos asignando a cada enlace un valor que refleja el número de caminos más cortos, que atraviesan ese enlace. Para otra medidas de centralidad que caractericen los enlaces se puede consultar a [6].

2.4. Cohesión de la red. Existen varias maneras de evaluar la cohesión de una red, dependiendo del problema, donde puede usarse triadas o componentes gigantes así como también lo que se denomina *cliques*, que no son más que subconjuntos de nodos totalmente cohesivos, en el sentido de que todos los vértices dentro del subconjunto están conectados por enlaces. Se pueden definir cliques de tamaño 1 (que en este caso son los nodos v) mientras que cliques de tamaño 2 representan los enlaces (e); los cliques o subgrafos de tamaño 3 son lo que también [13] denomina *triangles*, de manera que al ir aumentado el tamaño de los cliques es posible observar es la estructura del grafo bajo análisis. Una consecuencia del proceso de construcción antes descrito es que al aumentar el tamaño del clique, los últimos contienen los niveles más bajos, por lo cual [13] definen un concepto de *clique máximo*,llamado *clique number*.

2.5. Conectividad. Una noción de conectividad es la que tiene que ver con el hecho de que si dado un subconjunto de k vértices (o enlaces) se quitan del grafo, el subgrafo restante aún permanece conectado. En particular un grafo $G(V,E)$ se llama *k-vértice-conectado* si el número de vértices $N_v > k$, y al eliminar cualquier subconjunto de vértices $X \in V$ de cardinalidad $|X| < k$,X deja un subgrafo conectado. A su vez si $G(V,E)$ se denomina *k-borde-conectado* si $N_v \leq 2$, y al eliminar cualquier subconjunto de aristas $Y \in E$ de cardinalidad $|Y| < K$ deja un subgrafo que está conectado.

De esa manera se define como *conectividad* de vértice (enlace) de $G(V,E)$ al mayor entero tal que G es *k-vértice- (k-borde-)* conectado. En [13] los autores manifiestan que se puede demostrar que la conectividad del vértice está acotada por la conectividad de enlace, la que a su vez está acotada por el grado mínimo d entre los vértices en G.

2.6. Clustering de la red. Cuando se habla de partición de la red $\mathcal{C} = \{C_1, C_2, ..., C_k\}$, de un conjunto $\mathcal{S}$ se refiere a la división de la misma en clases naturales tales que estas son disjuntas entre sí y a su vez la unión de ellas reproducen el conjunto de partida ($\bigcup_{k=1}^{K} C_k = S$). Pero a su vez es importante también evaluar si un subconjunto de nodos (algunas de esas clases) es 'cohesivo' para lo cual se entiende que es así si los nodos están bien conectados entre sí, y al mismo tiempo están relativamente bien separados de los nodos restantes. Así los algoritmos de particionado buscan una partición $\mathcal{C} =$

$\{C_1, C_2, ..., C_k\}$ de un grafo $G = (V, E)$, de manera que los conjuntos $E(C_k, C_{k'})$ de enlaces conectando nodos de C_k en $C_{k'}$ sea relativamente pequeña en comparación al conjunto $E(C_k)$ o $E(C_{k'})$ de enlaces que conectan nodos al interior de C_k.

Una primera forma de evaluar el particionado de la red es a través de clustering jerárquico, de tipo aglomerativo, donde se incorpora una función de costo, que refleja la cohesión, con lo cual surge el concepto de *modularidad* de $\mathcal{C}$, donde se define $f_{ij}(\mathcal{C})$ como la fracción de enlaces de la red original que conectan nodos de C_i con nodos de C_j

$$mod(C) = \sum_{k=1}^{K} [f_{kk}(C) - f_{kk}^*]^2, \tag{4}$$

donde f_{kk}^* es el valor esperado de f bajo el supuesto de un modelo aleatorio de asignación de enlaces. Valores grandes de la *modularidad* sugieren que $\mathcal{C}$ captura una estructura *no trivial de grupos* (es decir que existen grupos), a la inversa que si los enlaces se asignasen al azar.

2.7. Enlace selectivo (Asortatividad).

Otro aspecto importante para evaluar la topología de una red es la evaluación de lo que se denomina *enlace selectivo entre nodos* de acuerdo a algunas características y que se miden con lo que se conoce como coeficiente de asortatividad (Assortativity coefficients) cuya lógica es muy similar a la de los coeficientes de correlación. Este concepto, que a veces se conoce como *homofilia*, expresa la tendencia de las personas a relacionarse con personas que se le parecen.

Cuando la característica que se estudia es de tipo categórico (nominal u ordinal) la medida es:

$$r_a = \frac{\sum_i f_{ii} - \sum_i f_{i+} f_{+i}}{1 - \sum_i f_{i+} f_{+i}} \tag{5}$$

Donde f_{ij} es la fracción de enlaces en $G(V, E)$ que unen un nodo en la i-ésima categoría con un nodo en la j-ésima categoría y f_{i+},f_{+i} expresan la suma de la i-ésima fila y columna respectivamente, de la matriz resultante $\mathbf{f}$ de frecuencias [16], [17].

El coeficiente descrito en la ecuación (5) está acotado en el intervalo $[-1, 1]$, de modo que si es cercano a 0, la mezcla de nodos en el grafo no difiere de la que se obtendría al asignar los enlaces al azar, preservando la distribución de grados marginal. Por otro lado, cuando el coeficiente se acerca a 1 o -1 existe una mezcla selectiva perfecta.

Cuando los nodos tienen una característica de interés que es continua, para evaluar la *homofilia* se consideran como (x_e, y_e) los valores que toman los nodos enlazados por el enlace e, para lo cual se usa el coeficiente de correlación de Pearson de los pares(x_e, y_e)

$$r = \frac{\sum_{x,y} xy - (f_{xy} - f_{x+} f_{+y})}{\sigma_x \sigma_y} \tag{6}$$

3. Descripción del problema en estudio

Para evaluar como funciona el AR, se trabaja con los datos provenientes del 'Relevamiento en población que se asiste Facultad de Odontología 2015 (RPAFO2015)', estudio sobre personas que demandan atención en la Facultad de Odontología de la Universidad de la República, Uruguay [1], donde se analiza el número de piezas cariadas (componente C), tratando de identificar su distribución, para luego estimar modelos de regresión. Este estudio se aplicó a una muestra de 602 personas que consultan en el período que corresponde a mayo 2015-junio 2016, los que se seleccionan mediante muestreo sistemático. Se les aplica un cuestionario sociodemográfico y un examen completo de la boca, en donde se evalúa el estado de las piezas dentales y de la mucosa, además de medidas antropométricas, de presión arterial y de glicemia. El tamaño muestral se determinó para estimar prevalencias de hasta 25 % con un margen de error $\delta = 0{,}05$ y un nivel de confianza $1 - \alpha = 0{,}95$ y cubrir hasta una tasa de respuesta del 90 %. Finalmente de los 640 originalmente calculados se obtuvieron 602, que representa una fracción de muestreo de alrededor del 10 % del total de personas que consultan anualmente.

En particular se consideran los siguientes atributos que conforman 3 bloques de variables:

Variable	Descripción	Bloque	Tipo
V1	Fuma a diario	1	Comportamental
V2	Consumo nocivo de alcohol	1	Comportamental
V3	Actividad física insuficiente	1	Comportamental
V4	IMC sobrepeso/obesidad,	2	ENT
V5	Razón de Cintura Cadera	2	ENT
V6	Hipertesión	2	ENT
V7	Diabetes	2	ENT
V8	Prev. bolsa	3	Odontológicas
V9	Pérdida Dentaria	3	Odontológicas
V10	Prevalencia de Caries	3	Odontológicas
V11	Prevalencia de PIP	3	Odontológicas

Tabla 1. Bloques de variables **ENT** utilizadas

Las primeras 3 variables constituyen factores de riesgo (bloque 1) que muchas veces se asocian con las variables del bloque 2 (**ENT**), que son a su vez patologías y que también son factores de riesgo a su vez para las variables del tercer bloque (patologías odontológicas).

[1] Pacientes evaluados por los odontólogos del Servicio de registros de la Facultad, desarrollado en el marco del proyecto 'Investigación y Desarrollo' de la Comisión Sectorial de Investigación Científica (CSIC), 2014 de la Universidad de la República

Variable	Descripción	Prevalencia
V1	Fuma a diario	**33.1**
V2	Consumo nocivo de alcohol	**9.8**
V3	Actividad física insuficiente	**44.7**
V4	IMC sobrepeso/obesidad	**57.3**
V5	Razón de cintura/cadera	**56.3**
V6	Hipertesión	**43.2**
V7	Diabetes	**21.3**
V8	Presencia bolsa	**58.6**
V9	Pérdida dentaria	**59.6**
V10	Prevalencia de caries	**72.8**
V11	Prevalencia de pip	**63.6**

Tabla 2. Prevalencias de variables estudiadas

4. Resultados

Para el análisis global se trabaja con el software [18] y para el análisis de los datos desde la perspectiva de redes se trabaja con la librería *igraph* [9].

A continuación se presenta el análisis de los datos mediante AR y para la detección de comunidades, se usa un conjunto de librerías [7], [9], [14].

Al trabajar con los datos desde la perspectiva del AR se construye un grafo, al cual se asocia la matriz de adyacencias y sobre la que se crean las comunidades mediante los algoritmos de *Random Walk* y *Fast Greedy*. La matriz de adyacencias surge de considerar los nodos como las 602 personas del **RPAFO2015**, y se establece que 2 nodos están conectados si comparten alguna de las 11 variables que se presentaron en el Cuadro 1. En las diferentes figuras la representación del grafo y la posición de los nodos es siempre la misma usando para eso un *layout* que es aleatorio pero donde se usa la misma semilla para inicializar la visualización de manera de tener siempre en el mismo lugar los mismos nodos y poder hacer comparables las figuras.

Los 601 individuos finalmente considerados, luego de remover el único individuo que quedaría aislado, generan un grafo conectado, el cual se describe en base a métricas como los grados y la frecuencia con las que se da cada patrón, así como medidas de centralidad. En la Tabla 2 se consigna las prevalencias de cada variable en forma global.

Puede verse en el Cuadro 3 que la configuración que más veces aparece es la que corresponde al nodo 206, con 12 nodos iguales que tiene un alto número de enlaces siendo un nodo tipo que tiene presente las 3 patologías del tipo **ENT** del bloque 2 y 2 de las 4 bucales. A este perfil de nodo se le antepone el nodo 97 (que es 1 de los 2 que aparece con esta configuración) y que está mucho menos conectado y se caracteriza por ser una persona que solamente presenta el ́habito de fumar. El resto de las filas del Cuadro 3 muestran otras configuraciones que corresponden a nodos con mayor número de enlaces (mayor grado), pero que aparecen von menor frecuencia, lo que luego se verifica al ser asignados a diferentes clusters. Una característica que es común a los 6 nodos es que todos los individuos presentan las 4 patologías bucales.

	nodo	patrón	grados	frecuencia
Perfiles por frecuencia				
	206	0-0-1-1-1-1-0-1-0-1-1	586	12
	97	1-0-0-0-0-0-0-0-0-0-0	198	2
Perfiles con mayor grado				
	205	1-0-1-0-1-1-0-1-1-1-1	600	13
	175	1-0-1-1-1-0-0-1-1-1-1	600	13
	281	1-0-1-1-1-1-0-1-1-1-1	600	13
	506	1-0-1-1-1-1-1-1-1-1-1	600	13
	324	1-1-1-1-1-0-0-1-1-1-1	600	13
	171	1-1-1-1-1-1-0-1-1-1-1	600	13

Tabla 3. Descripción de algunos nodos

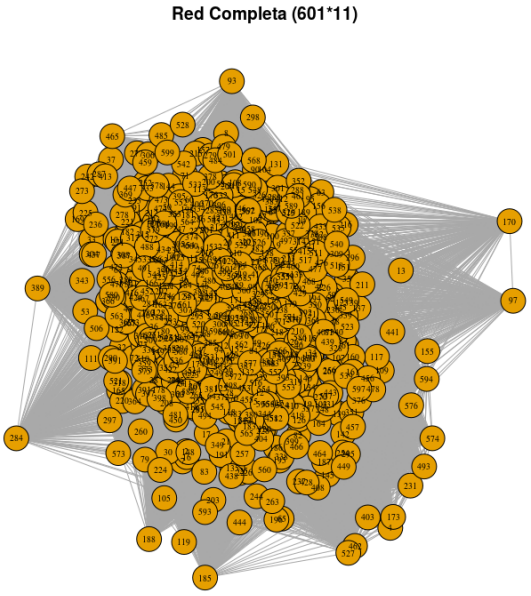

Fig. 2. Red generada con 601 individuos analizados

Sobre el grafo se aplican los dos algoritmos de detección de comunidades ya mencionados, donde la modularidad es mayor para la solución de 2 grupos que surge del algoritmo.

5. Discusión

Con respecto a la caracterización de los grupos que surgen de las comunidades detectadas con el **AR** se puede comentar los siguientes hallazgos (con respecto al método de Fast Greedy):

- Grupo 1, con un total de (n=276) individuos que muestran mayor prevalencia de las variables

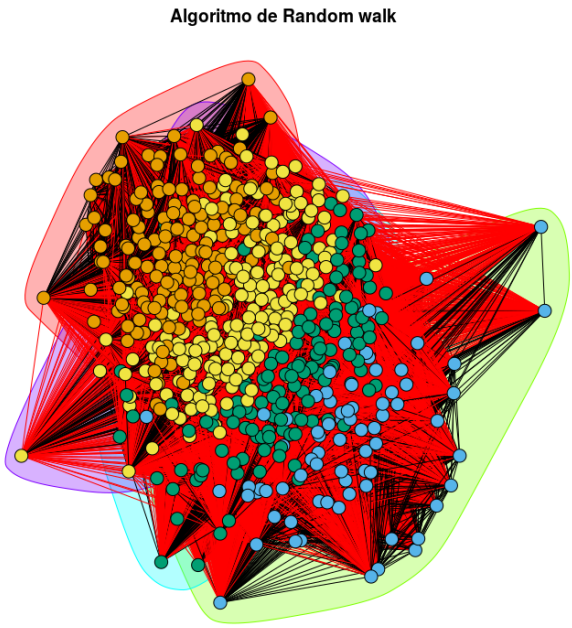

Fig. 3. Comunidades identificadas con Algoritmo Random Walk

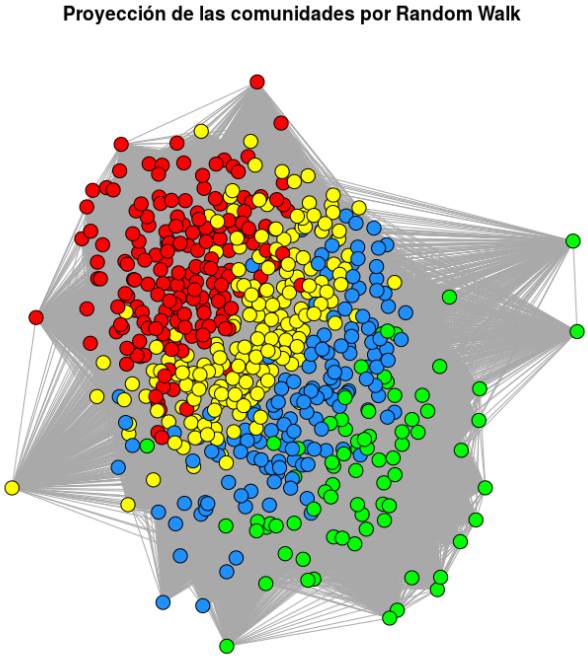

Fig. 4. Proyección en la red de los grupos encontrados con algoritmo Random Walk

Algoritmo Cluster Random Walk	Algoritmo (Cluster Fast Greedy) 1	2	Total
1	0	177	177
2	76	0	76
3	146	3	149
4	54	145	199
Total	276	325	601

Tabla 4. Comparación de los 2 métodos de detección de comunidades

	(Método Random walk)				(Método Fast Greedy)		
Variable	1	2	3	4	1	2	Total
V1	14.7	44.7	44.5	36.1	50.7	18.1	**33.1**
V2	7.9	9.2	12.1	10.0	10.8	8.9	**9.8**
V3	55.3	27.6	38.9	46.2	38.0	50.4	**44.7**
V4	91.5	0.00	10.0	84.4	18.8	90.1	**57.3**
V5	96.0	0.00	4.0	81.9	15.6	91.8	**56.3**
V6	67.2	5.3	29.5	47.7	25.4	58.4	**43.2**
V7	35.6	0.00	9.4	25.6	9.0	31.6	**21.3**
V8	58.7	21.0	78.5	58.3	60.1	57.5	**58.6**
V9	7.8	94.7	71.8	83.4	81.1	41.5	**59.6**
V10	59.9	71.0	78.5	80.9	80.4	66.6	**72.8**
V11	80.1	3.9	79.8	59.8	55.8	70.5	**63.6**
Tamaño	**177**	**76**	**149**	**199**	**276**	**325**	

Tabla 5. Perfiles de los grupos creados mediante AR

comportamentales (lo que supone un aumento de los factores de riesgo salvo para la actividad física insuficiente), mientras que presenta una disminución de la prevalencia de las variables ENT. En cuanto a las variables odontológicas, se observa un perfil con mayor carga de patología.

- Grupo 2, con mayor número de individuos (n=325), presenta una disminución del consumo de tabaco y de alcohol, mientras que la prevalencia de actividad física insuficiente está por encima del promedio. A su vez es un grupo caracterizado por tener una alteración de las ENT muy por encima de los valores medios, mientras que en cuanto a las variables odontológicas se observa un comportamiento de menor prevalencia.

Las Figuras 3, 4 presentan la partición de la red resultante luego de aplicar el algoritmo *Random Walk*. Pueden observarse las comunidades detectadas, claramente diferenciadas, donde el grupo 1 es el rojo, el 2 el verde, el 3 el azul y el 4 el amarillo. A continuación se presenta la caracterización de estos cuatro grupos.

- Grupo 1, con un total de (n=177) individuos, presenta mayor prevalencia de las variables ENT, con un gran incremento en las variables metabólicas (IMC y Razón de Cintura). En cuanto a las variables del

tercer bloque, se observa un perfil con mayor carga de patología odontológica, en particular la pérdida dentaria y la prevalencia de Pérdida de inserción periodontal.

- Grupo 2, con la menor cantidad de individuos (n=76), presenta un aumento del consumo de tabaco, casi la media global de consumo nocivo de alcohol y una prevalencia de actividad física insuficiente por debajo del promedio en casi un $40\,\%$. A su vez, es un grupo caracterizado por alteraciones de las ENT muy por debajo de los valores medios y un comportamiento en cuanto a las variables odontológicas de menor prevalencia salvo para pérdida dentaria, casi duplicando el valor medio.
- Grupo 3, con mayor cantidad de individuos que el grupo 2 (n=149), caracterizado por mayot carga de factores comportamentales (tabaco y alcohol) por encima de la media, con poca carga de los factores de riesgo del bloque ENT, pero con una fuerte carga en cuanto a la patología bucal, con un aumento de casi $50\,\%$ con respecto a la media global.
- Grupo 4, el más numeroso (n=199), que podría resumirse como un grupo con una elevada carga de factores comportamentales, factores de riesgo metabólicos y con mucha mas patología bucal, salvo para la pérdida de inserción periodontal.

Como todo problema de clustering, este no escapa a la situación donde no se sabe exactamente el número de grupos, sino mas bien aproximaciones a un nuúmero óptimo. No es de extrañar que una vez clasificados los individuos, las comunidades presenten perfiles que se alejan de los centroides de éstas a pesar de estar más próximos de éstas que si estuviesen en otros grupos. Es posible mejorar esta situación generalizando el método a un método mixto mediante un proceso de difusión donde cada individuo puede extenderse a otros grupos con diferentes grados de membresía, partiendo previamente de una clasificación previa, que podría ser la resultante de los algoritmos de detección de comunidades para AR.

6. Conclusiones

Con los resultados encontrados hasta el momento, aplicando la metodología de AR,se aprecia una partición en comunidades estable.

A su vez, tal como se decía en la introducción, el objetivo era plantear una alternativa de elaboración de perfiles mediante una estrategia metodológica particular, pero que consiste en no tener en cuenta la jerarquía que existe en las variables al considerarlas a todas en igualdad de condiciones para segmentar la población bajo estudio. La literatura muestra que en general la aproximación es de tipo modelizante donde hay claramente un bloque de variables explicadas, que serían las del bloque de variables odontológicas, las que pueden ser explicadas por las variables comportamentales (bloque 1), siendo estas las que configuran factores de riesgos, mientras que las variables ENT (pretenecientes al bloque 2) que indican patologías y a su vez factores de riesgo de las odontológicas. Por eso se proponen varios caminos que puedan ayudar a entender mejor las tipologías resultantes y la de detección de comunidades mediante AR.

- Determinar las comunidades dentro de cada bloque de variables y cruzar entre sí las particiones elaboradas, comparando con los resultados que surgen al considerar las once variables sin agruparlas jerárquicamente.
- Crear la tipología con la lógica de clustering convencional (clasificación no supervisada) usando por ejemplo un método basado en el algoritmo *k-modes* que es de tipo *modal*, y que no es más que un caso particular de un un *k-prototipo* descrito por [12]. En este caso, el algoritmo tiene una lógica de funcionamiento similar a la del algoritmo *k-means*, y dada la naturaleza de las variables, es necesario el uso de otras medidas de disimilaridad, usando un método basado en frecuencias para actualizar los modos [23].
- Comparar los resultados con los comunidades detectadas con el AR, como se hizo en este trabajo
- Proponer un análisis de redes validando modelos estadísticos (recordar que esto es solo descripción), donde algunos de los atributos evaluados en la caracterización se pueden usar como variables explicativas, usando la teoría de los modelos exponenciales aleatorios en grafos (ERGM), [13].

Referencias

[1] Bhupathiraju SN, T. K. (2011). Coronary heart disease prevention: Nutrients, foods, and dietary patterns. *Clin Chim Acta*, 412(17-18):1493–514.

[2] Bonacich, P. (1987). Power and centrality: A family of measures. *American Journal of Sociology*, 5:1170.

[3] Bonacich, P. and Lloyd, P. (2001). Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, 23:191 – 201.

[4] Borgatti, S. P., Everett, M. G., and Johnson, J. (2013). *Analyzing Social Networks*. SAGE Publications Ltd.

[5] Brandes, U. (2001). A faster algorithm for betweenness centrality. *The Journal of Mathematical Sociology*, 25(2):163–177.

[6] Brandes, U. and Erlebach, T. (2005). *Network analysis: methodological foundations*. Number 3418 in LCNS, Tutorial. Springer, Berlin ; New York. OCLC: ocm58474176.

[7] Butts, C. T. (2016). *sna: Tools for Social Network Analysis*. R package version 2.4.

[8] Cook, N., Cutler, J., Obarzanek, E., Buring, J., Rexrode, K., and SK, K. (2007). Long term effects of dietary sodium reduction on cardiovascular disease outcomes: observational follow-up of the trials of hypertension prevention (tohp). *British Medical Journal*, 334(7599):885–8.

[9] Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

[10] Food and Agriculture Organization of the United Nations (2010). Fats and fatty acids in human nutrition. Report of an experte consultation 10-14 novem- ber 2008, FAO.

[11] Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215.

[12] Huang, Z. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. in kdd: Techniques and applications. Technical report, World Scientific.

[13] Kolaczyk, E. and Csárdi, G. (2014). *Statistical analysis of network data with R*. Springer, New York.

[14] Kolaczyk, E. and Csárdi, G. (2017). *sand: Statistical Analysis of Network Data with R*. R package version 1.0.3.

[15] Luke, D. (2015). *A user's guide to network analysis in R*. Springer.

[16] Newman, M. E. J. (2002). Assortative mixing in networks. *Phys. Rev. Lett.*, 89:208701.

[17] Newman, M. E. J. (2003). Mixing patterns in networks. *Phys. Rev. E*, 67:026126.

[18] R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[19] R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[20] Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31(4):581–603.

[21] Skapino, E. y Alvarez Vaz, R. (2016). Prevalencia de factores de riesgo de enfermedades crónicas no transmisibles en funcionarios de una institución bancaria del Uruguay. *Revista Uruguaya de Cardiología*, 31:246 – 255.

[22] Wasserman, S. and Faust, K. (1994). *Social network analysis: methods and applications*. Number 8 in Structural analysis in the social sciences. Cambridge University Press, Cambridge ; New York.

[23] Weihs, C., Ligges, U., Luebke, K., and Raabe, N. (2005). klar analyzing german business cycles. In Baier, D., Decker, R., and Schmidt-Thieme, L., editors, *Data Analysis and Decision Support*, pages 335–343, Berlin. Springer-Verlag.

[24] World Cancer Research Fund International (2007). Food, nutrition, physical activity and the prevention of cancer: a global perspective. Technical report, World Cancer Research Fund, American Institute for Cancer Research.