El análisis reticular de coincidencias

Networked analysis of coincidences

Modesto Escobar
Carlos Tejero
Universidad de Salamanca
modesto@usal.es (ESPAÑA)

Recibido: 03.12 2016
Aceptado: 19.12.2017

RESUMEN

El propósito de este artículo es la propuesta de un nuevo marco para el estudio de las estructuras de datos basado en la combinación de diversos análisis multivariantes y de redes sociales. Mediante estas técnicas se obtienen qué sucesos son los más frecuentes en un conjunto de escenarios y con qué otros sucesos tienden a ocurrir.

A este respecto pueden distinguirse diversos gradientes de coincidencias entre los sucesos estudiados, que van desde la nula coincidencia hasta la coincidencia total pasando por las coincidencias estadísticamente probables con nivel de confianza prefijado. La estructura de aparición del conjunto de los sucesos estudiados conforme al gradiente de coincidencia seleccionado puede ser convenientemente representada mediante un grafo.

Además de sus fundamentos, se presentan tres programas gratuitos con los que cualquier usuario podría aplicarlo: coin, netcoin y webcoin.

Este tipo de procedimiento puede ser aplicado al análisis exploratorio de cuestionarios, al estudio de redes semánticas, a la revisión de bases de datos e incluso a la comparación de distintas técnicas de análisis estadísticos de interdependencia, al hacer uso de métodos factoriales, clasificatorios y distintos modelos de representación de grafos basados en fuerzas de atracción-repulsión como los de Fruchterman-Reingold y Kamada-Kawai.

PALABRAS CLAVE

Análisis estadístico, análisis de redes sociales, estadística visual interactiva, coincidencias, grafos.

ABSTRACT

The goal of this paper is the proposal of a new framework for the study of data structures based on the combination of several types of multivariate and social network analysis. By means of these techniques we obtain the most frequent events in a given set of scenarios as well as those events that tend to occur with them.

In this respect we can define several coincidence gradients for the events under study. Ranging from zero to total coincidence and passing through statistically probable coincidences with predetermined confidence levels.

The appearance structure of the set of events studied according to the selected coincidence gradient can be conveniently represented by a graph.

In addition to its rationale, three free software programs are shown so that any user could apply this framework: coin, netcoin and webcoin.

This type of procedure can be applied to the exploratory analysis of questionnaires, to the study of semantic networks, to the revision of databases and even to the comparison of different techniques of statistical analysis of interdependence. This is made posible by using factorial and classificatory methods and different methods for representing graphs based on attraction-repulsion forces, like those of Fruchterman-Reingold and Kamada-Kawai

KEY WORDS

Statistical analysis, Social networks analysis, Interactiv visual statistics, coincidences, graphs.

1. INTRODUCCIÓN

Breiman (2001) escribió que existen dos grandes culturas en el uso de la estadística para obtener conclusiones a partir de los datos. Por un lado, se encuentra la *cultura del modelaje*, que asume como punto de partida un modelo de datos estocásticos, a modo de caja negra que conecta los predictores (o variables independientes) con los resultados (o variables dependientes). Es el caso de los económetras, que emplean como punto de partida el modelo lineal o alguna de sus generalizaciones, como son los modelos logísticos, multinomial, o binomial negativo (Gujarati y Dawn 2008, Green 2012, Wooldridge 2016). Por otro lado, se ha desarrollado una *cultura algorítmica*, que considera el interior de esa supuesta caja negra, como algo complejo y desconocido, y en consecuencia, la tarea de la estadística consiste en encontrar la función que conecta los predictores con las respuestas. Bajo estos supuestos, se encuentra toda la tradición del aprendizaje automático (*machine learning*), especializado en el uso de las redes neuronales, los árboles de clasificación, o las máquinas de soporte vectorial. En ambas tradiciones, se emplea la estadística como una herramienta capaz de des-

cubrir los determinantes de los fenómenos estudiados (Hastie et al. 2009, James et al. 2013, Zhao, 2013). Sin embargo, Breiman olvidó que habría una tercera vía o *cultura* en el tratamiento de los datos, que podríamos denominar la multivariante, o mejor aún, estadística *de la interdependencia*, ya que no distingue entre predictores y respuestas. Antes, al contrario, trata toda la información o conjunto de variables con el mismo status y trata de descubrir las regularidades existentes en los datos. En este sentido, los análisis factoriales, las escalas multidimensionales, los análisis de conglomerados o incluso los modelos log-lineales son todas ellas herramientas que se emplean comúnmente en la literatura científica para encontrar pautas de distribución en conjuntos amplios de información (Kendall 1975, Everitt and Dunn 2001, Afifi et al. 2012).

Aunque esta tercera cultura no tenga un alto prestigio en la comunidad científica, es indudable su necesidad en el análisis de datos, pues puede cumplir al menos tres funciones. En primer lugar, juega un importante papel en el análisis exploratorio (Tukey 1977) si se desea ir más allá del examen de sólo una o dos variables al mismo tiempo. Cuando se dispone de mucha información, es conveniente aplicar una primera aproximación gráfica a nuestros datos. Para ello, además de los gráficos de cajas, los de estrellas o las caras de Chernoff (1973) son capaces de mostrar las diferentes configuraciones que adoptan las variables en distintos casos. En segundo lugar, este tipo de técnicas ha sido y viene siendo usado frecuentemente para la simplificación de la información. A este respecto, ¿quién no ha empleado un análisis de componentes principales para convertir un gran número de variables en una cantidad reducida de factores? o ¿quién no ha intentado aplicar a una muestra un análisis de conglomerados a fin de clasificar los casos en un número pequeño de grupos? En tercer lugar, los análisis multivariantes juegan un papel muy importante en la minería de datos, especialmente, en el campo de los inmensos conjuntos de datos (big data) de que se dispone en la actualidad. Aunque en este terreno han destacado los portadores de la segunda cultura, es decir, la algorítmica; el análisis multivariante también puede contribuir a descifrar y hacer inteligible la gran cantidad de información que está al alcance del ciudadano en la sociedad de la información y de las nuevas tecnologías.

Además de estas tres culturas estadísticas, matizadas o no con la perspectiva bayesiana (Carlin y Louis 2000, Bernardo y Smith 2000), en las últimas décadas se han desarrollado otros tipos de análisis de los fenómenos. Aquí destacaremos dos de ellos: el primero tiene un nombre confuso, pues ha sido denominado bajo la etiqueta de teoría de juegos, cuando en realidad (Myerson 1991) trata del estudio de los modelos matemáticos de conflicto y cooperación entre decisores racionales. El segundo, que será incorporado de modo central en esta presentación, es el análisis de redes sociales, consistente en el estudio de la estructura de los vínculos entre personas u organizaciones, así como las consecuencias que de ella se desprenden (Wasserman y Faust 1994, Scott 2017).

El propósito de este artículo consiste en ofrecer un nuevo modelo para la representación de la estructura de los datos. No se pretende generar una nueva cultura estadística, ni tan siquiera descubrir algoritmos revolucionarios para el

tratamiento de grandes cantidades de datos; sino más bien la integración de los ya existentes y desarrollar una serie de herramientas para extender su uso en el campo de las ciencias sociales.

A la presente propuesta se la denominará Análisis Reticular de Coincidencias (ARC), puesto que su principal objetivo es descubrir una serie de fenómenos, opiniones o características que en un determinado campo suelen aparecer conjuntamente. Muchos estadísticos persiguen la ilusión de dar con las causas de los fenómenos a partir de la información. Sin embargo, hay que ser cautos pues, salvo que se aplique con rigor el método experimental, las herramientas estadísticas son muy limitadas en el estudio de causas y efectos. Por ello, se propone una serie de análisis que no tienen como meta el descubrimiento de las "verdaderas" causas de los fenómenos en estudio, sino sus pautas de concurrencia con el fin de proporcionar al investigador posibles sugerencias de cómo está estructurada la realidad.

Hay una amplia variedad de instrumentos en el campo del análisis de datos que desempeñan objetivos similares a los que se plantean con este modelo. En primer lugar, ha de mencionarse el análisis de datos exploratorios (Tukey 1977; Escobar 1999), que tiene un importante papel en el primer acercamiento a la información disponible. Del mismo modo, cabe encontrar una gran similitud con ciertas técnicas encuadradas dentro del aprendizaje automático (Witten et al. 2005 y Flach 2012) como son las reglas de asociación (Agrawal et al. 1993), que buscan asociaciones entre eventos de orden dos y superior. Por su lado, el análisis cualitativo comparado (Ragin 1987, 2000; Medina et al. 2017) comparte los datos binarios como entrada, aunque use un procedimiento distinto de tratamiento de la información basado en la lógica de (Bool 2003). Finalmente, cabe referirse al llamado análisis de coocurrencias que se aplica básicamente en dos ámbitos: en el de las estructuras comunitarias de especies (Sanderson 2000, Griffith et al. 2016) y en el del análisis de contenido basado en redes semánticas (Dagan et al. 1999, Matsuo e Ishizuka 2002), que se centra en el número de veces que aparecen determinados vocablos en un conjunto determinado de unidades de texto.

Para alcanzar el objetivo de este artículo, se presentará en primer lugar los fundamentos estadísticos en los que está basado el llamado análisis reticular de coincidencias. A continuación, se dará cuenta de tres aplicaciones gratuitas desarrolladas para llevar a cabo este tipo de análisis. Finalmente se presentarán algunos ejemplos de su posible uso.

2. DEFINICIONES DE COINCIDENCIAS

Definición 1: Un *experimento aleatorio* es un procedimiento con resultado imprevisto que puede repetirse indefinidamente.

Definición 2: Cada resultado potencial de un experimento aleatorio es llamado un *suceso* (*j*). El conjunto de posibles resultados se denomina espacio muestral y está compuesto por una serie de *sucesos elementales* mutuamente

excluyentes.

Definición 3: Un *escenario* (*i)* es cada uno de los resultados de un experimento complejo compuesto por un conjunto de sucesos (X_j) con mayor o menor grado de dependencia entre sí. También puede considerarse escenario un delimitado conjunto espacial y temporal en el que el investigador recoge información sobre los sucesos que en aquél tienen lugar. Como los sucesos de los escenarios no son mutuamente excluyentes, conviene representarlos por vectores dicotómicos (se presentan o no en el escenario) o naturales (número de veces que ocurren en un determinado escenario).

Por tanto, el conjunto de escenarios observados puede ser representado como una *matriz de incidencias* (**I**). En ella se recogen en una dimensión, generalmente la de las filas, los escenarios (*I*) y en la otra dimensión, comúnmente la de las columnas, los sucesos (*J*). Esta matriz se compone únicamente de 0 y 1, indicando respectivamente la ausencia o presencia de los sucesos en los escenarios. Como alternativa, sin embargo, puede trabajarse con la *matriz de ocurrencias*, en la que puede registrarse la aparición de más de un suceso de la misma clase en el mismo escenario.

Con un sencillo ejemplo, se comprenderá mejor esta distinción. Imaginemos que tenemos cuatro escenarios en cada uno de los cuales se lanzan dos monedas y estamos interesados en los sucesos cara y cruz. Las tres posibilidades de resultados serían: a) dos caras, ninguna cruz; b) una cara con una cruz, y c) dos cruces con ninguna cara. Una matriz de ocurrencias podría presentarse del siguiente modo:

Tabla 1.- Matriz de ocurrencias

	Sucesos	
Escenarios	Cara	Cruz
I	2	0
II	1	1
III	1	1
IV	0	2

En cambio, en una matriz de incidencias, solo se reflejan los valores 0 y 1, según los valores de ocurrencia sean 0 o mayor que 0:

Tabla 2.- Matriz de Incidencias

	Sucesos	
Escenarios	Cara	Cruz
I	1	0
II	1	1
III	1	1
IV	0	1

A partir de las matrices de incidencias y ocurrencias pueden obtenerse las respectivas matrices de coincidencias y coocurrencias.

Definición 4: Dos sucesos (*j* y *k*) reciben la calificación de *coincidentes* si ocurren conjuntamente en el mismo escenario *i*.

$$(x_{ij} = 1 \wedge x_{ik} = 1) \Rightarrow f_{ijk} = 1$$

Además de la coincidencia elemental en un escenario *i*, al estudiar si dos eventos son coincidentes en un conjunto múltiple de escenarios pueden distinguirse distintos grados de coincidencia. $f_{jk} = \sum_{i=1}^{I} f_{ijk}$. De este modo, la más elemental clasificación de coincidencias distinguiría entre:

a) *Coincidencia nula*: Dos sucesos nunca aparecen en el mismo escenario $(f_{jk} = 0)$, diciéndose, por tanto, que son *mutuamente excluyentes*.

b) *Coincidencia simple*: Dos sucesos son meramente coincidentes si aparecen conjuntamente en al menos un escenario $(f_{jk} > 0)$.

c) *Coincidencia total*: Dos sucesos aparecen siempre conjuntamente en los mismos escenarios. Si uno de ellos aparece en un escenario, *necesariamente* ocurre el otro $(f_{jk} = f_{jj} = f_{kk})$. Un caso especial de este tipo de coincidencias sería la *coincidencia subtotal*, que implica que solo ocurre el otro suceso si aparece el primero y no viceversa $(f_{jk} = f_{jj} < f_{kk})$, esto es, que la aparición del suceso más frecuente (*k*) no necesariamente implique la aparición del suceso (*j*) menos frecuente.

Para estudiar el comportamiento de las coincidencias puede obtenerse su matriz mediante la siguiente expresión: $\mathbf{F} = \mathbf{I'I}$. Los elementos de esta matriz son tanto frecuencias univariadas (f_{jj}) como bivariadas (f_{jk}) de los distintos sucesos en el conjunto de escenarios (I) que se expresan en las filas de $\mathbf{I}$.

De las matrices $\mathbf{F}$ de frecuencias, pueden derivarse tres medidas probabilísticas: las probabilidades marginales, las condicionales y las probabilidades conjuntas.

La *probabilidad marginal* de X_j, denotada como $\Pr(X_j)$, puede obtenerse a partir del cociente entre las frecuencias de cada suceso (f_{jj}) y el número total de escenarios donde podía haber aparecido (I).

$$\Pr(X_j) = \frac{f_{jj}}{I}$$

La *probabilidad conjunta* de dos sucesos X_j y X_k, expresada como $\Pr(X_{jk})$ viene dada por la frecuencia de que ocurran en el mismo escenario divida también por el conjunto de escenarios contemplados en un determinado conjunto:

$$\Pr(X_{jk}) = \frac{f_{jk}}{I}$$

Las probabilidades *condicionadas*, a las que denotaremos por $\Pr(X_j \mid X_k)$, expresan la posibilidad de que haya ocurrido un determinado suceso, en el supuesto de que haya ocurrido otro segundo suceso. Se obtienen dividiendo la probabilidad conjunta y la probabilidad marginal del suceso condicionante.

$$\Pr(X_j \mid X_k) = \frac{\Pr(X_{jk})}{\Pr(X_k)} = \frac{f_{jk}}{f_{kk}}$$

A partir del concepto de probabilidad condicionada puede considerarse el gradiente de *coincidencia probable* entre dos sucesos en el caso de que la proba-

bilidad del primero condicionada por la ocurrencia del segundo[1] sea mayor del 50%.

$$\Pr(X_j \mid X_k) > 0,5$$

Como en muchas ocasiones se trabaja con muestras de escenarios, en lugar del universo de ellos, puede estimarse el límite inferior de su intervalo de confianza bajo la hipótesis alternativa de que $\Pr(X_j \mid X_k) < .50$ mediante la fórmula

$$L_{\text{inf.}} = \frac{f_{jk}}{f_{kk}} - \frac{t_{\alpha, f_{kk}-1}}{2\sqrt{f_{kk}}}$$

siendo $t_{\alpha, f_{kk}-1}$ el valor de la distribución de Student para $f_{kk} - 1$ grados de libertad con un nivel de significación α.

Otro gradiente de coincidencia es el de *coincidencia condicional* o *dependiente*. Se obtiene a partir del concepto de independencia de sucesos. Dos sucesos son independientes si se da la siguiente igualdad:

$$\Pr(X_j) = \Pr(X_j \mid X_k) \Leftrightarrow \frac{f_{jj}}{I} = \frac{f_{jk}}{f_{kk}}$$

Por tanto, para que se cumpla esta condición, ha de verificarse la siguiente condición:

$$f_{jk} = \frac{f_{jj} f_{kk}}{I}$$

A partir de esta igualdad, dos sucesos dados tienen coincidencia dependiente siempre que su frecuencia sea mayor que la esperada (f^*_{jk}) bajo el supuesto de independencia.

$$f_{jk} > \frac{f_{jj} f_{kk}}{I} = f^*_{jk}$$

Además se sabe (Haberman, 1968) que la diferencia entre f_{jk} y f^*_{jk} asume asintóticamente una distribución normal con el siguiente error típico:

$$\sqrt{f^*_{jk}(1 - f_{jj} / I)(1 - f_{kk} / I)}$$

Por esta razón, los residuos pueden normalizarse, empleando la fórmula del residuo de Haberman (r_{jk}):

$$r_{jk} = \frac{(f_{jk} - f^*_{jk})}{\sqrt{f^*_{jk}(1 - f_{jj} / I)(1 - f_{kk} / I)}}$$

Resumiendo este apartado, los distintos grados de coincidencias que pueden detectarse entre cada par de sucesos son los siguientes:

[1] Nótese que el concepto de coincidencia probable es asimétrico, lo cual quiere decir que el suceso A puede ser probable con respecto el B sin implicar esto que el suceso B lo sea con respecto al A. En general, $\Pr(X_j \mid X_k) \neq \Pr(X_k \mid X_j)$.

Tabla 3.- Tipos de coincidencias

Tipo de coincidencia	Definición	Asimétrica	Prueba estadística
Nula.....	$f_{jk} = 0$	No	No
Simple	$f_{jk} > 0$	No	No
Probable	$f_{jk} / f_{kk} > 0,5$	Sí	Sí
Dependiente	$f_{jk} > f^*_{jk}$	No	Sí
Subtotal	$f_{jk} = f_{jj} < f_{kk}$	Sí	No
Total	$f_{jk} = f_{jj} = f_{kk}$	No	No

3. MEDIDAS DE COINCIDENCIAS

Además de clasificarlas en distintos tipos, las coincidencias pueden ser medidas empleando para ello las medidas de proximidad binaria (Hubálek, 1982; Gower, 1985). Estas medidas poseen un valor máximo de uno cuando hay total coincidencia entre dos sucesos dicotómicos y 0 cuando hay total independencia entre ellos. Algunas de ellas, pueden adoptar valores negativos, en cuyo caso el valor mínimo podría ser -1, en el caso de dos sucesos completamente antagónicos, es decir, cuando uno aparece, el otro no está presente y viceversa.

Para el cálculo de estas medidas puede partirse de cada casilla (f_{jk}) de la matriz de coincidencias con el siguiente sistema de equivalencias:

$$a = f_{jk}$$
$$b = f_{jj} - f_{jk}$$
$$c = f_{kk} - f_{jk}$$
$$d = I - f_{jj} - f_{kk} + f_{jk}$$

Por tanto, para cada par de sucesos, puede elaborarse una tabla bidimensional de frecuencias del siguiente aspecto:

Tabla 4.- Nomenclatura de los sucesos según su presencia o ausencia

	Suceso X_k	
Suceso X_j	Presente	Ausente
Presente	a	b
Ausente	c	d

Con estas cuatro cantidades que representan las frecuencias de los cuatro estados de presencia/ausencia de dos sucesos en el conjunto de escenarios estudiados pueden obtenerse las llamadas medidas de proximidad binaria (Hubálek 1982).

Estos coeficientes o medidas de proximidad binaria pueden clasificarse en cuatro tipos. En el primero, se incluyen las medidas similares a la de *matching* (también conocida como la de *Rogers y Tanimoto*), pues son un cociente entre un numerador en el que aparecen tanto las coincidencias positivas (los dos sucesos aparecen en el mismo escenario), como las coincidencias negativas (los dos sucesos están ausentes en el mismo escenario), y un denominador en el que se contemplan todos los escenarios, aunque con distinto peso. Otras medidas que pertenecen a esta categoría son la de *Rogers*, la de *Sneath*, la de *Anderberg* y la de *Gower2*.

$$Rogers\ y\ Tanimoto = \frac{a+d}{a+b+c+d}$$

$$Rogers = \frac{a+d}{(a+d)+2(b+c)}$$

$$Sneath = \frac{2(a+d)}{2(a+d)+(b+c)}$$

$$Anderberg = \left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{c+d}+\frac{d}{b+d}\right)/4$$

$$Gower = \frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$$

El segundo tipo de medidas son las similares a las de *Jaccard*. En ellas, se excluyen aquellos escenarios en los que no aparece ninguno de los dos sucesos cuyo nivel de coincidencia se pretende medir. Por tanto, no se incluyen ni en el numerador ni en el denominador aquellos escenarios sin ninguno de los dos sucesos. Medidas con el mismo criterio serían las de *Dice*, *Antidice*, *Ochiai* y *Kulczynski*.

$$Jaccard = \frac{a}{a+b+c}$$

$$Dice = \frac{2a}{2a+b+c}$$

$$Antidice = \frac{a}{2+2(b+c)}$$

$$Ochiai = \frac{a}{\sqrt{(a+b)(a+c)}}$$

$$Kulczynski = \left(\frac{a}{a+b}+\frac{a}{a+c}\right)/2$$

En el tercer tipo de medidas de similitud para datos binarios, solo podría catalogarse la de *Rusell*. Esta se caracteriza por considerar semejantes solo los escenarios en los que aparecen ambos eventos (a). De este modo, los eventos coincidentes por ausencia en los mismos escenarios son excluidos en el numerador, como ocurría en las medidas del tipo anterior. En contraste, a diferencia de las medidas similares a *Jaccard*, aparecen todos los escenarios posibles (*a+b+c+d*) en el denominador.

$$Rusell = \frac{a}{a+b+c+d}$$

Finalmente, en el cuarto apartado pueden incluirse todas aquellas medidas en las que en el numerador se comparan (sustraen) las frecuencias de coincidencias (tanto si aparecen como si no aparecen los fenómenos) con las frecuencias de no coincidencias (escenarios en los que aparece un fenómeno, pero está ausente el otro). Como consecuencia, estas mediciones pueden ser positivas, si predominan sucesos coincidentes, o negativas, en caso contrario, es decir, cuando predominan los escenarios en los que los sucesos no coinciden. A este apartado pertenecen las conocidas medidas de *Hamman* y *Yule*, así como la de *Pearson*.

$$Hamann = \frac{(a+d)-(b+c)}{a+b+c+d}$$

$$Yule = \frac{ad-bc}{ad+bc}$$

$$Pearson = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

Ahora bien, en su primer cálculo todas estas medidas son llamadas de similitud. Para convertirlas en medidas de distancia, podemos proceder transformándolas de acuerdo a la expresión distancia=1-similitud. Si la medida de similitud tiene un rango entre 0 y 1, estos límites se conservan en la correspondiente distancia, pero con un significado diferente, pues el 0 indicará ahora completa coincidencia. Bastará con convertirla, con la fórmula acabada de mencionar. En el caso de que sea una medida con rango entre -1 y +1, la nueva medida de similitud estará comprendida entre 0 y 2, indicando el 1 completa independencia y los coeficientes mayores que esta cantidad son expresión de que dos sucesos coinciden con menor frecuencia que la que implicaría el mero azar.

4. LAS MATRICES DE ADYACENCIAS

Se ha hablado de matrices de coincidencias y de similitud con su contrapartida, las matrices de distancias. Tanto unas como otras podrían convertirse en matrices de adyacencias. Se define como tal una matriz que conecta o no cada par de sucesos en función del valor que presentan en una medida de coincidencia. Por tanto, es una matriz cuadrada con tantas filas y columnas como sucesos se estén estudiando y formadas por elementos binarios que representan la coincidencia o no de dos sucesos. En función de todas las medidas vistas con anterioridad pueden formarse matrices de adyacencias de los siguientes modos:

a) Con las coincidencias *simples* de tal modo que haya una conexión entre dos sucesos con tal de que coincidan en un solo escenario.

b) Con las coincidencias *totales* o *subtotales* de forma que solo se conecten dos sucesos completamente coincidentes. En la primera categoría la conexión será simétrica, no siendo así en el caso de coincidencias subtotales, en las que solo habrá vínculo entre la categoría menos frecuente y la más frecuente.

c) Con las coincidencias *probables* o *condicionales*, conectando los sucesos con más del 50% de probabilidad en el primer caso y con un residuo (r_{jk}) positivo en el segundo.

d) Con las *pruebas estadísticas* aplicadas a las coincidencias *probables* o *condicionales*, en cuyo caso podríamos disponer de coincidencias estadísticamente significativas en distintos grados o niveles de significación (0,05, 0,01, 0,001, 0,0001…)

e) Con las *medidas de similitud*, en cuyo caso ha de optarse por una de las 14 posibles, estableciendo un umbral —0,50, por ejemplo— a partir del cual pueda considerarse que dos sucesos son coincidentes.

Tabala 5.- Diversos estadísticos de coincidencias con matrices de adyacencias

Coincidencias				Adyacencias							
Coincidencias	1	I	P	Coin. simple	1	I	P	Coin. Matching	1	I	P
Resultado =1	**19**			Resultado =1				Resultado =1			
Resultado impar	19	**53**		Resultado impar	1			Resultado impar	1		
Resultado <=3	19	34	**50**	Resultado <=3	1	1		Resultado <=3	1	1	
Pr. condicional	1	I	P	Coin. total	1	I	P	Coin. Jaccard	1	I	P
Resultado =1		0.36	0.38	Resultado =1				Resultado =1			
Resultado impar	1.00		0.68	Resultado impar	0			Resultado impar	0		
Resultado <=3	1.00	0.64		Resultado <=3	0	0		Resultado <=3	0	0	
Pr. residuos	1	I	P	Coin. subtotal	1	I	P	Coin. Dice	1	I	P
Resultado =1				Resultado =1		1	1	Resultado =1			
Resultado impar	0.00			Resultado impar	0		0	Resultado impar	1		
Resultado <=3	0.00	0.00		Resultado <=3	0	0		Resultado <=3	1	1	
Jaccard	1	I	P	Coin. probable	1	I	P	Coin. Rusell	1	I	P
Resultado =1				Resultado =1		0	0	Resultado =1			
Resultado impar	0.19			Resultado impar	1		1	Resultado impar	0		
Resultado <=3	0.19	0.34		Resultado <=3	1	1		Resultado <=3	0	0	
Yule	1	I	P	Coin. condicional	1	I	P	Coin. Yule	1	I	P
Resultado =1				Resultado =1				Sacar un 1			
Resultado impar	1.00			Resultado impar	1			Resultado impar	1		
Resultado <=3	1.00	0.55		Resultado <=3	1	1		Resultado <=3	1	1	

Fuente: Experimento de 100 lanzamientos de un dado de 6 caras.
Abreviaturas: 1: resultado del dado igual a 1; I: resultado impar; P: resultado <=3.
N.B. Las matrices triangulares son simétricas. Salvo en la matriz de coincidencias, la diagonal de la matriz se omite por considerarse irrelevante.

En la tabla 5 aparecen 10 matrices de adyacencias en las dos columnas de la derecha. En la de la izquierda se han expuesto algunas de las matrices de similitud de las que proceden: Desde la de coincidencias hasta la matriz de los coeficientes de Yule. Como puede observarse los resultados difieren según se emplee uno u otro criterio. Hay plenas coincidencias simples, probables y condicionales. De igual modo, mediante las medidas de *matching*, Dice y Yule los tres sucesos pueden considerarse coincidentes, pero no lo son, si se tienen en cuenta las medidas de Jaccard o Rusell. Sacar un 1, implica sacar un número impar o un número pequeño; pero no a la inversa. Por tanto, ambos pares de sucesos pueden ser catalogados como coincidencias subtotales.

5. GRÁFICOS DE COINCIDENCIAS

Hay diversos modos de representar gráficamente las coincidencias. Empezaremos con los más sencillos, que son los diagramas de barras. Entre estos, cabe distinguir, en primer lugar, los gráficos de *barras de incidencias* que sirven solo para diferenciar la correspondiente frecuencia de cada suceso en los escenarios estudiados. Mediante estos gráficos puede compararse la distinta incidencia que presentan los eventos o sucesos en términos del porcentaje de escenarios en los que se encuentran presentes.

El segundo tipo de *gráficos de barras* es el que podríamos denominar gráfico *de coincidencias*. En este caso, cada suceso ha de aparecer en un gráfico específico junto con el conjunto restante de sucesos mediante una simple barra para cada uno de ellos con dos tonalidades: una más larga que representa la incidencia del suceso de la barra y otra tonalidad más corta con frecuencia proporcional a la coincidencia de este suceso con el del gráfico. De este modo, se puede apreciar fácilmente con qué otros sucesos tiene más coincidencias el suceso representado.

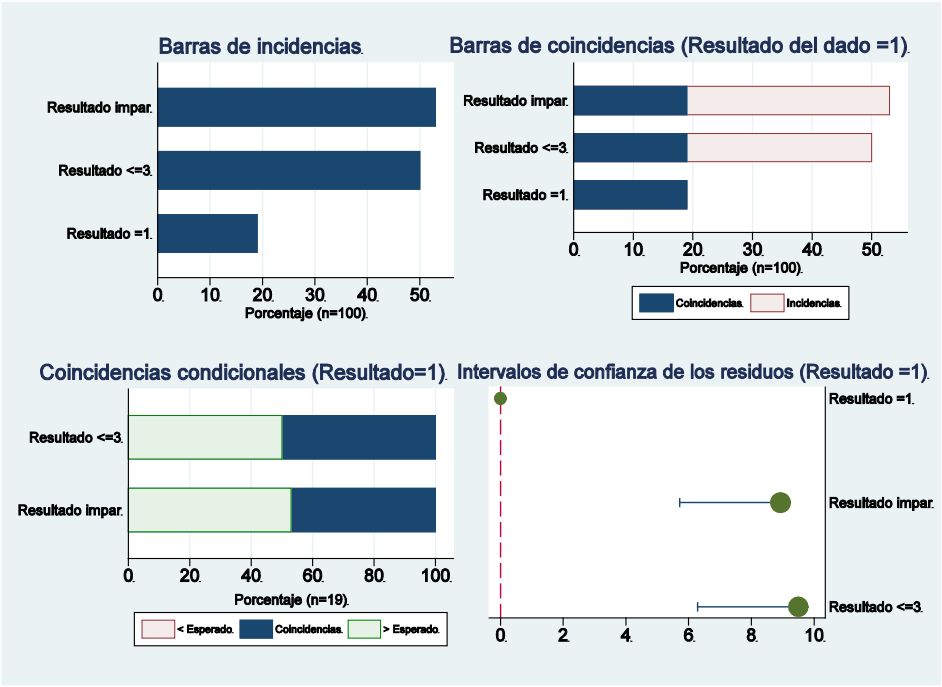

Figura 1.- Gráficos de coincidencias obtenidos tras el lanzamiento de un dado 100 veces.

El tercer tipo de *gráfico de barras* es el *de* las *coincidencias condicionales*. Al igual que el anterior, se centra sobre la relación de un suceso con todos los restantes representados cada uno de ellos mediante una barra. Ahora bien, en lugar de tener en cuenta todos los escenarios, solo aparecen reflejados en el gráfico aquellos en los que ha aparecido el suceso representado. Las barras indican, como en el tipo anterior, porcentajes, pero en esta ocasión los porcentajes son condicionales, es decir, el cociente entre la frecuencia de aparición conjunta y la frecuencia del suceso en consideración. En cada barra, además de representar el porcentaje condicional de aparición del suceso de dicha barra sobre el suceso del gráfico, también se representa el porcentaje del suceso condicionado con el fin de saber si la probabilidad condicional es mayor o menor que éste. En el caso de que sea menor (aparece a la izquierda), los dos sucesos serán condicionalmente coincidentes. En el caso opuesto de que sea mayor, entonces no podrían hablarse de coincidencia. En el gráfico situado abajo a la derecha de la figura 4, con tono más oscuro están representados los porcentajes condicionales de las coincidencias (en los dos sucesos representados adoptan un valor del 100%, puesto que todos los unos son menores que tres e impares. Con tono más claro, se dibujan las frecuencias relativas de las incidencias de sacar un número impar y de sacar un resultado bajo. Como ambas están por debajo de los valores anteriores, representados con barra oscura, quiere decir que hay coincidencia condicional. Además,

como ambas también alcanzan el 100%, se trata de coincidencias subtotales.

De naturaleza distinta son los gráficos de residuos (y de cociente de razones). En ellos también se representa la coincidencia de un suceso frente al resto de los posibles sucesos que se encuentran en los escenarios. Pero, en lugar de expresar las frecuencias relativas, se da cuenta de los residuos estandarizados, es decir, de las diferencias entre las frecuencias empíricas y las frecuencias esperadas en el supuesto de que los sucesos fuesen independientes. El valor del residuo se expresa en la escala horizontal del gráfico y de él parte el intervalo de confianza de una sola cola dirigido hacia el valor de la hipótesis nula, es decir, el valor 0.

De similar hechura sería el gráfico de cociente de razones. En este sentido, los valores de estos estadísticos variarían entre 0 e infinito, representando el valor unitario la independencia entre los sucesos. Tanto más a la derecha se representaran los sucesos, tanto más coincidentes serían con el suceso principal que encabeza el gráfico.

6. GRAFOS DE COINCIDENCIAS

Ahora bien, de todos los tipos de gráficos que sirven para estudiar coincidencias el que más información puede ofrecer es el grafo de coincidencias. Consiste este en dibujar todos los sucesos que interesan u ocurren en los escenarios como nodos o vértices vinculados entre sí por aristas, siempre y cuando sean coincidentes.

Estos grafos[2] pueden experimentar importantes variaciones con el fin de mejorar la representación de los fenómenos estudiados. A continuación, se exponen las principales: el tamaño, el color, la forma y la posición de los nodos.

El tamaño juega un papel muy importante en la información que pueden trasmitir los grafos de coincidencias. El tamaño puede afectar tanto a los nodos, que representan los sucesos, como a las aristas, que dan cuenta de las coincidencias entre ellos.

En el primer caso, lo más común en los grafos es que el tamaño sea proporcional al grado de cada nodo (número de conexiones con otros nodos). Sin embargo, en un grafo de coincidencias, se considera que lo que puede proporcionar mejor información es el nivel de incidencia de cada suceso de modo tal que sean tanto mayores cuanto con más frecuencia aparezcan en los escenarios.

En el segundo caso, la tradición en los grafos de coocurrencias es dibujar los vínculos o aristas con un grosor proporcional a la frecuencia conjunta de los dos sucesos vinculados. No obstante, para las coincidencias es preferible emplear un doble criterio: primero, no representando los vínculos que no superen determinado umbral (coincidencias probables, condicionales aplicando preferiblemente en ambos casos criterios de significación estadística); segundo, empleando como peso del grosor de los vínculos una medida de la coincidencia (residuos de Ha-

[2] Se encuentran ejemplos de estos grafos en las figuras 1-3 del último apartado del artículo.

berman, Pearson o los estadísticos de significación de estos, entre los preferidos).

En lo que se refiere al color y la forma, se suelen emplear para la categorización de los sucesos o nodos. Por ejemplo, en el caso de los resultados de lanzar un dado, se podrían colorear los sucesos compuestos y dejar en blanco los elementales (o viceversa). De igual modo, se podría emplear un cuadrado para resaltar el suceso sacar un número pequeño con el fin de resaltarlo frente al resto de sucesos, que podrían dibujarse, como es habitual, con círculos. Las categorías con las que se configuran color y forma no tienen por qué ser conocidas de antemano. Con ayuda de otras herramientas estadísticas (análisis de conglomerados, por ejemplo) también podrían clasificarse los sucesos estudiados, incluso en función de sus coincidencias entre ellos.

Finalmente, también merece una atención especial la posición de los sucesos en el grafo. A tal fin, pueden distinguirse dos tipos de algoritmos: los basados en modelos físicos de atracción/repulsión y los que se sustentan en modelos estadísticos de naturaleza principalmente factorial.

En el primer apartado, los más conocidos y empleados son el de Fruchterman y Reingold (1991), por un lado, y el de Kamada-Kawai (1989), por otro. En el segundo apartado, el más versátil y conocido es el que proviene de las escalas multidimensionales (Cox y Cox 2001), pero también podrían aplicarse otros sistemas de asignar las coordenadas a los sucesos (o nodos) en función de criterios del análisis factorial (Everitt 2003: 296), el de correspondencias (Everitt 2003: 94) o los biplots (Gabriel 1971).

7. PROGRAMAS PARA REPRESENTAR LAS COINCIDENCIAS

Ahora bien, para realizar el análisis mencionado, hasta ahora era necesario emplear distintos programas y conectar los resultados de unos con los otros. Así, en un análisis de contenido se comenzaba con Atlas-ti, un programa de tratamiento cualitativo de datos, para codificarlos. El resultado se convertía a SPSS para tabularlos; luego se pasaba a Excel para calcular los residuos de Haberman y, finalmente, se realizaban los análisis de redes con Ucinet, Pajek o Netdraw. Recientemente, a fin de que el proceso se difunda y se emplee por el mayor número posible de investigadores e investigadoras, se están elaborando programas capaces de realizar todo el proceso sin necesidad de cambiar de aplicación, al menos a partir de la matriz de datos o de la codificación aplicada a los datos. Ello se ha realizado con Stata; pero también se está realizando en otros sistemas más extendidos, como puede ser R o Java Script. A continuación, se exponen brevemente algunos de estos programas.

coin[3]

coin es un programa de usuario elaborado con Stata que genera tanto estadís-

[3] El programa puede ser descargado gratuitamente desde Stata mediante la siguiente instrucción dentro de la aplicación: net describe st0416, from(http://www.stata-journal.com/software/sj15-4)

ticas como gráficos a partir de un conjunto de variables dicotómicas dispuestas en un conjunto de datos (Escobar, 2015). En este sentido, los casos son considerados como escenarios y las variables como sucesos. Opcionalmente también trabaja con un segundo conjunto de datos en el que los registros son características de los sucesos que pueden incluirse para seleccionarlos o representarlos de modo distinto.

En el terreno numérico *coin* es capaz de generar tablas de frecuencias de las coincidencias, porcentajes condicionales (de modo horizontal o vertical), porcentajes conjuntos, frecuencias esperadas en caso de coincidencias independientes, residuos brutos, estandarizados y normalizados, así como las matrices de adyacencias y medidas de distancia y centralidad de los sucesos.

En el campo gráfico este programa produce gráficos de incidencias, de coincidencias simples y condicionales, así como de residuos y cociente de razones. También es capaz de representar dendrogramas y grafos de coincidencia de los sucesos con posiciones circulares, escalas multidimensionales, componentes principales, correspondencias y biplot, así como es capaz de emplear también el algoritmo de Fruchterman-Reingold.

webcoin[4]

webcoin es un programa desarrollado con R-Shiny capaz de representar mediante un grafo las coincidencias existentes en un fichero que el usuario de una página web sube desde su propio ordenador.

Este fichero puede estar en formato csv (con encabezados y comas como separadores entre campos), SPSS o Stata. Sus registros o casos deberán ser los escenarios. El primer campo o variable deberá ser el nombre del escenario, mientras que los restantes campos o variables serán los sucesos expresados como variables dicotómicas.

Una vez subido el fichero, el usuario será capaz de ver el contenido del fichero (los 50 primeros escenarios) en la pestaña *Table*, un gráfico con nodos estáticos en la pestaña *Plot* y otro dinámico en la llamda *D.Graph*.

En ambos tipos de gráficos se puede controlar el tipo de coincidencia. Para ello, los grados expuestos son: coincidencia simple, condicional, condicional significativa (.05), condicional bastante significativa (.01) y condicional muy significativa (.001) De igual modo la mínima proporción de incidencias que ha de tener un suceso (nodo) para ser representado en el grafo también puede variarse a través de uno de los controles situados en el margen izquierdo de la aplicación.

De modo menos crucial también puede cambiarse el tamaño relativo de los nodos y las aristas. Los grosores de estos son respectivamente proporcionales a la incidencia de los sucesos y a los residuos normalizados de Haberman.

En el grafo estático puede cambiarse la disposición (*layout*) de los nodos, habiendo para ello distintas modalidades: Fruchterman-Reingold, Kamada-Kawai, escalamiento multidimensional, círculo, estrella y aleatoria. También se pueden representar mediante colores y sectores los distintos tipos de sucesos según las

[4] Para usar este programa ha de accederse a la siguiente dirección: http://coin.der.usal.es:8080/Upload/

comunidades o bloques que los conforman.

Finalmente, en el grafo dinámico, representado siempre con el algoritmo de Fruchterman-Reingold, puede variarse la carga de repulsión entre los nodos no conectados entre sí. Valores más negativos implican mayor distancia entre los sucesos en el caso de que no sean coincidentes.

netcoin[5]

netcoin es una librería escrita en R que permite al usuario generar matrices de coincidencias con sus correspondientes grafos y crear páginas web interactivas a partir de ellas.

En dichas páginas interactivas pueden cambiarse un gran conjunto de elementos de los grafos, así como generar tablas y gráficos descargables.

Entre los elementos modificables se citan los siguientes:

a) La etiqueta, el tamaño, el color y la forma de los sucesos o nodos, en función de sus propiedades. De igual modo se pueden representar áreas de nodos con las mismas características (conglomerados) e incluso reemplazar las formas geométricas de los nodos con imágenes.

b) La etiqueta, el grosor y el color de las aristas que representan las coincidencias entre los sucesos, en función de las propiedades de los vínculos (frecuencias, grado de coincidencia, significación…)

c) Se puede hacer una selección de nodos manual o en función de sus atributos.

d) Se permite realizar una selección de las aristas en función de sus propiedades.

La disposición de los nodos del grafo puede asumir dos modalidades. De entrada, las coordenadas pueden determinarse por el usuario del paquete. Al pulsar el botón correspondiente, los nodos se ubican en una disposición basada en el criterio de Fruchterman-Reingold, cuyas fuerzas de atracción y repulsión pueden ser cambiadas.

Con los nodos seleccionados se forman dos tipos de tablas de atributos: la de los sucesos y la de coincidencias. En cada una de estas tablas aparecen las correspondientes propiedades de unos y otras, características que el usuario de la página web puede descargarse en su propio ordenador.

8. USOS Y EJEMPLOS

El primer uso que se le puede dar a este análisis proviene de la dificultad de trabajar en cuestionarios con preguntas multirespuesta. Por ejemplo, una pregunta tan elemental como los medios que los parados han empleado para encontrar empleo. Para esta pregunta, los cuestionarios en uso incluyen una serie de posibles respuestas como “a través de familiares”, “a través de conocidos”,

[5] Un ejemplo de aplicación de este paquete puede verse en http://coin.der.usal.es/CIS. Al ser de dominio público en el repositorio CRAN, se puede descargar una versión de este paquete mediante la siguiente instrucción de R: install.packages(«netCoin»).

"mandando CV a las empresas", "a través de anuncios en los periódicos", "en las oficinas del INEM" … La primera dificultad de este tipo de preguntas de cuestionario es su codificación, ya que no basta una sola columna para esta tarea, como en el caso de las preguntas de alternativas múltiples. Existen dos formas de realizarlo: una dejando tantas columnas como el máximo de respuestas posibles. Piénsese en que se le preguntara sobre los tres medios más empleados. En ese caso, no son necesarias más que tres variables (o columnas) con valores mutuamente excluyentes. Sin embargo, si se deja libre el número de respuestas, pueden llegar a ser necesarias tantas columnas, que podrían codificarse bien de modo múltiple (del 1 al 5 por haber cinco opciones) o bien de forma dicotómica, dejando cada columna para una de las opciones de búsqueda de trabajo y codificando con uno aquellos casos en los que se mencionara la opción y un cero (o un blanco) si no lo hicieran.

El problema se complica más en el análisis, pues no resulta fácil hacer cruces combinadas con variables de esta naturaleza. Hasta que no se escribieron tareas específicas llamadas *mult-response*, los análisis se tenían que hacer por separado. Incluso, en las primeras versiones de estas rutinas resultaba imposible cruzar una variable múltiple consigo misma, para ver en una sola tabla las frecuencias de los métodos de búsqueda de empleo coincidentes. Es más, hoy en día, aunque sea posible cruzar una variable múltiple consigo misma, poco más que los porcentajes pueden ser obtenidos, a menos que se hagan tantos cruces como pares de valores múltiples se dispongan.

Mediante cualquiera de las aplicaciones mencionadas en el apartado anterior es posible realizar los cálculos de los residuos estandarizados y presentarlos conjuntamente en una tabla como la 6 o en un formato de gráfico reticular.

Tabla 6.- Matriz de residuos estandarizados con las opciones de búsqueda de empleo por parados.

Residuos estandarizados	O.1	O.2	O.3	O.4	O.5	O.6	O.7	O.8	O.9	O.10	O.11	O.12	O.13
Oficina													
1.- INEM	31.6												
2.- Otra	**4.5**	31.6											
Contactos													
3.- Empresarios	1.2	**5.3**	31.6										
4.- Otros	**2.7**	**2.5**	**7.5**	31.6									
Anuncio													
5.- Colocando	0.3	**6.7**	**6.9**	**3.3**	31.6								
6.- Mirando	1.3	**5.6**	**8.1**	**5.2**	**16.6**	31.6							
Auto													
7.- Empleo	-1.2	-0.3	-2.5	-2.2	-0.5	-0.7	31.6						
8.- Préstamo	-0.3	1.3	-1.8	-1.8	1.2	0.4	**24.9**	31.6					
Esperando													
9.- Ofertas	**7.8**	**3.0**	**1.7**	**3.6**	**1.9**	1.4	-0.8	-1.0	31.6				
10.- Resultados	1.4	**4.3**	**4.9**	**2.7**	**7.6**	6.0	1.3	1.6	**9.0**	31.6			
Otros													
11.- Entrevistas	1.2	**4.9**	**2.6**	0.4	5.0	**3.8**	-0.2	0.6	**3.4**	**10.4**	31.6		
12.- Exámenes	-0.7	0.7	-0.6	-1.2	0.9	0.0	**2.0**	**2.1**	0.6	2.0	**4.6**	31.6	
13.- Otros	0.4	0.0	1.1	-1.0	1.0	-0.1	1.2	1.3	**1.7**	-0.2	1.1	-0.3	31.6

En negrita los residuos estandarizados con valores positivos significativos ($p<0,05$) de una cola.
Fuente: EPA. Segundo trimestre 2013. INE.

El siguiente uso del análisis reticular de coincidencias es el análisis de contenido. En un estudio para la fundación ECOTEC, Quintanilla et al. (2011) analizaron los libros de texto españoles de la ESO, a fin de descubrir la cultura científica que estos materiales trasmiten. Para ello, se escanearon 81 textos de cuatro editoriales y se realizó con la ayuda de Atlas.ti una codificación de los principales conceptos científicos. De ellos, se seleccionaron 227 para el análisis por ser que tuvieron más de 200 apariciones en los 134.397 párrafos analizados. Finalmente, para ver cómo se articulaban a través de los textos, se emplearon los residuos ajustados de Haberman (véase infra), para entresacar aquellos que aparecían con más regularidad en los mismos párrafos y, de esta manera, no solo se presentaba la frecuencia de aparición, sino también con qué otros conceptos aparecían. A tales efectos, se empleó el análisis de redes a partir de una matriz de adyacencias construida bajo la suposición de que dos conceptos están vinculados si aparecen de modo significativo en los mismos párrafos. Empleando estas técnicas se elaboraron gráficos como el presentado en la figura 1.

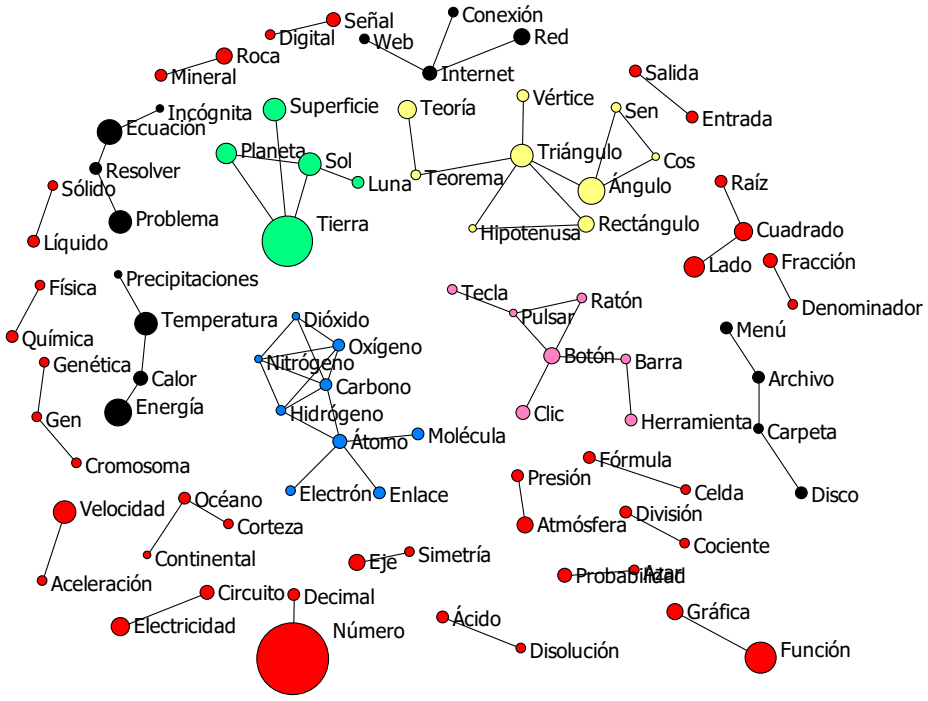

Figura 1.- Red semántica de la ciencia en manuales de la ESO

Otro uso similar de esta técnica se empleó para analizar el contenido de textos periodísticos (Escobar, 2009). En concreto, se empleó el archivo hemerográfico del Profesor Juan Linz sobre la transición española en la prensa

(1973-1987), que había sido catalogado por la Biblioteca del Centro de Estudios Avanzados en Ciencias Sociales del Instituto Juan March con un extenso tesauro de 10.000 descriptores. Entre estos, 442 aparecieron en más de 200 artículos diferentes y fueron analizados con el ARC, obteniendo, entre otros, el siguiente gráfico que resume los contenidos más relevantes, así como su conexión entre ellos (figura 2).

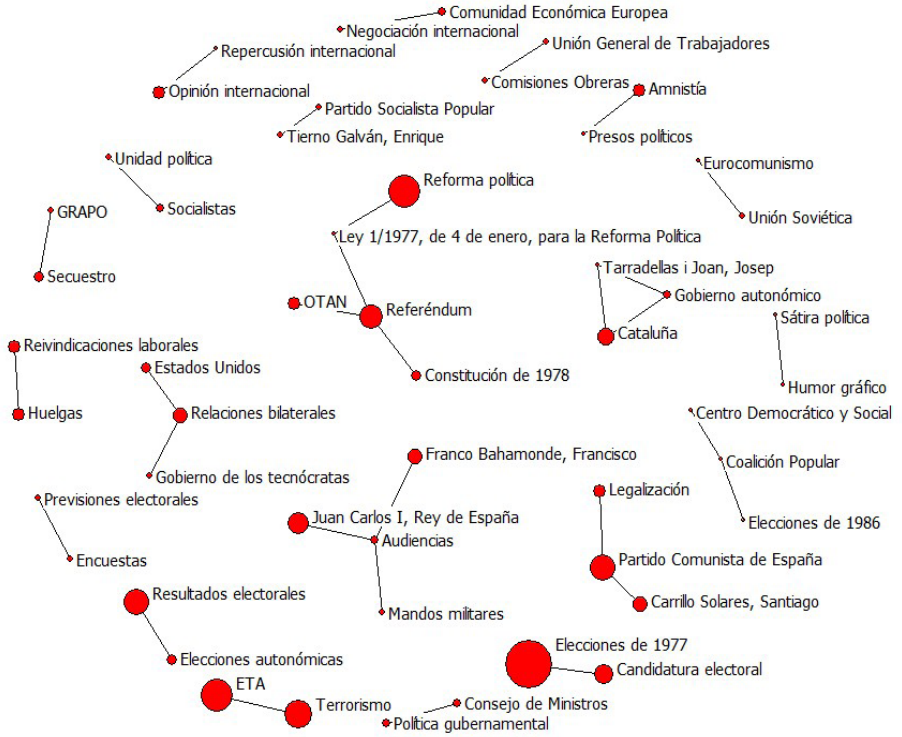

Figura 2.- Red semántica de actores y temas de la transición española

También este análisis ha sido empleado para analizar documentos fotográficos (Escobar, 2015). En concreto, se rescataron dos álbumes familiares de principios del siglo XX con el objetivo de estudiar qué personajes aparecen en las mismas fotografías: el de Miguel de Unamuno y el del músico sevillano Joaquín Turina. Para analizar el primero, se empleó el fichero de catalogación de las fotos elaborado por la Casa de Unamuno de la Universidad de Salamanca, convirtiendo el campo de los 27 personajes que aparecían en más de 2 fotografías en otras tantas 27 variables dicotómicas. Fruto del ARC, se dibujó una red en la que el centro era el filósofo bilbaíno, que aparecía junto a otros personajes en las

fotografías de sus viajes etnográficos, al tiempo que también estaba fuertemente vinculado con los miembros de su familia, esposa, hijos y nietos, que a su vez estaban estrechamente conectados, especialmente sus ocho hijos con su madre. (Véase figura 3).

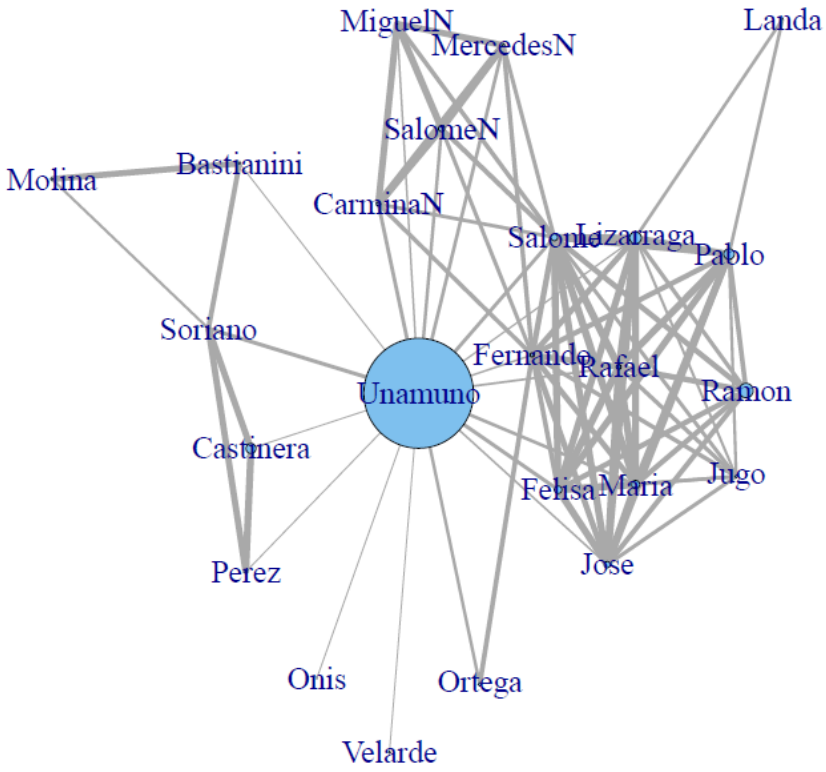

Figura 3.- Grafo de las coincidencias de personas en la colección fotográfica de Miguel de Unamuno

9. CONCLUSIONES

El análisis reticular de coincidencias tiene como finalidad descubrir las pautas de concurrencia de una serie de sucesos en un conjunto de escenarios. Su objetivo es descubrir cómo se distribuyen conjuntamente una serie de características dispuestas en distintas unidades en las que pueden o no estar presentes.

Pueden distinguirse diversos grados de coincidencias: la nula, la simple, la probable, la dependiente, la estadísticamente probable y dependiente, la subtotal y la total. Además, para estimar qué grado de coincidencias presentan dos suce-

sos puede emplearse un amplio rango de estadísticos: desde la mera frecuencia, hasta medidas de distancias más complejas, como la de Haberman o la de Rusell. Los programas *coin, webcoin*, y *netcoin* permiten obtenerlas haciendo uso de Stata o R. Muchos de estos estadísticos también pueden ser obtenidos con otras aplicaciones como SPSS, SAS, Excel…

Ahora bien, no solo es importante representar las coincidencias numéricamente con estadísticos. Una buena representación gráfica puede ayudar sobremanera a entender mejor la distribución de las coincidencias de un conjunto múltiple de sucesos. Mediante *coin*, se pueden representar las coincidencias con siete tipos de gráficos resumibles en cuatro modalidades:

Los gráficos de barras. Su principal ventaja es la sencillez, pero plantean el inconveniente de que generan un gráfico por cada suceso, por lo que serían necesarios muchos de ellos caso de que se quisiera abordar el estudio de las coincidencias de un conjunto no reducido de sucesos.

Gráficos de residuos, cuyo objetivo es la distinción entre sucesos probables y estadísticamente probables, por dibujar los intervalos de confianza de los residuos normalizados o de los cocientes de razones.

Los dendrogramas conjuntan todas las coincidencias en un solo gráfico; pero presentan el inconveniente de que los modos de aglomeración, así como las múltiples medidas de distancia disponibles, pueden distorsionar el estudio conjunto de los pares de coincidencias posibles.

Finalmente, mediante *coin*, *webcoin* y *netcoin* pueden obtenerse gráficos de redes que permiten representar no solo las incidencias de los fenómenos, sino también sus grados de coincidencias e incluso las características de los sucesos en consideración. Aunque haya distintos modos de ubicarlos en el espacio bidimensional de las representaciones gráficas, todos ellos aportan información similar y no contradictoria a través de los vínculos entre sucesos. De todos modos, se recomienda la representación mediante algoritmos basados en fuerzas o mediante escalas multidimensionales por la simplicidad de las distancias geodésicas que emplean y por la presentación de los sucesos más coincidentes en el centro del grafo.

10. BIBLIOGRAFÍA

AFIFI, A.A., S. MAY y V.A. CLARK (2012): Practical Multivariate Analysis, Boca Raton, CRC Press.

AGRAWAL, R., T. IMILINSKI y A. SWAMI (1993): Mining Association Rules between Sets of Items in Large Databases. Ponencia presentada en SIGMOD '93 Proceedings of the 1993 ACM SIGMOD International Conference on Management of data, Washington.

BERNARDO, J.M. y A.F.M. SMITH (2000): Bayesian Theory. Chichester, UK, Wiley.

BOOL, G. (2003): An Investigation of the Laws of Thought, Amherst, NY, Prometheus book.

BREIMAN, L. (2001): "Statistical Modeling: the Two Cultures", Statistical Science, 16(3), 199-131.

CARLIN, B.P. y T.A. LOUIS (2000): Bayes and Empirical Bayes Methods for Data Analysis, Boca Raton, FL, Chapman & Hall/CRC.
CHERNOFF, H. (1973): "The Use of Faces to Represent Points in K-Dimensional Space Graphically", Journal of the American Statistical Association*,* 68(342), 361-368.
DAGAN, I., L. LEE y F. PEREIRA (1999): "Similarity-Based Models of Word Co-occurrence Probabilities", Machine Learning*,* 34(1), 43-69.
ESCOBAR, M. (1999): Análisis gráfico/exploratorio, Madrid, La Muralla.
ESCOBAR, M. (2009): "Redes semánticas en textos periodísticos: propuestas técnicas para su representación", Empiria, 17, pp. 13-39.
ESCOBAR, M. (2015): "Studying coincidences with network analysis and other multivariate tools", The Stata Journal, 15(4), pp. 1118-1156.
ESCOBAR, M. y J. GÓMEZ ISLA (2015): "La expresión de la identidad a través de la imagen: los archivos fotográficos de Miguel de Unamuno y Joaquín Turina", Revista Española de Investigaciones Sociológicas, 152, pp. 23-46.
EVERITT, B. S. (2003): The Cambridge Dictionary of Statistics, Cambridge, Cambridge University Press.
EVERITT, B.S. y G. DUNN (2001): Applied Multivariate Data Analysis, London, Arnold.
FLACH, P. (2012): Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge, Cambridge University Press.
FRUCHTERMAN, T.M.J. y E.M. REINGOLD (1991): "Graph Drawing by Force-Directed Placement", Software: Practice and Experience, 21(11), 1129-1164.
GABRIEL, K. R. (1971): "The Biplot Graphical Display of Matrices with Application to Principal Component Analysis", Biometrika, 58, 453-467.
GOWER, J.C. (1985): "Measures of similarity, dissimilarity and distance", Encyclopedia of Statistical Sciences Vol. 5, New York, Wiley, pp. 397-405.
GREEN, W. H. (2012): Econometric Analysis, Boston, Pearson.
GRIFFITH, D. M., J.A. VEECH y C. MARSH (2016): "Cooccur: Probabilistic Species Co-occurrence Analysis in R", Journal of Statistical Software, 69(2), 1-17.
GUJARATI, D.N. y D.C. PORTER (2008): Basic Econometrics. Boston, McGraw Hill.
HABERMAN, E. (1973): "The Analysis of Residuals in Cross-classified Tables", Biometrics, 29, pp. 205-220.
HASTIE, T., R. TIBSHIRANI y J. FRIEDMAN (2009): The Elements of Statistical Learning, New York, Springer.
HUBÁLEK, Z. (1982): "Coefficients of Association and Similarity, Based on Binary (presence-absence) Data: An Evaluation", Biological Reviews, 578, 669-689.
JAMES, G. *et al*. (2013): An Introduction to Statistical Learning, New York, Springer.
KAMADA, T. Y S. KAWAI (1989): "An Algorithm for Drawing General Undirected Graphs", Information Processing Letters, 31(1), 7-15.
KENDALL, M. G. (1975): Multivariate Analysis, London, Griffin.
MATSUO, Y. Y M. ISHIZUKA (2002): "Keyword Extraction from a Document Using Word Co-Occurrence Statistical Information", Transactions of the Japanese Society for Artificial Intelligence, 17(3), 217-223.
MEDINA, I. *et al*. (2017): Análisis cualitativo comparado, Madrid, CIS.
MYERSON, R. B. (1991): Game Theory: Analysis of Conflict, Cambridge (Mass.), Harvard University Press.
RAGIN, C. C. (1987): The Comparative Method. Berkeley, University of California Press.
RAGIN, C. C. (2000): Fuzzy-Set Social Science. Chicago, University of Chicago Press.

SANDERSON, J. (2000): "Testing Ecological Patterns. A Well-known Algorithm from Computer Science Aids the Evaluation of Species Distributions", American Scientist, 88, 332-339.
SCOTTT, J. (2017): Social Network Analysis, Los Ángeles, Sage.
QUINTANILLA, M. Á., ESCOBAR, M., ESCRIBANO, M., & SABBATINI, M. (2005). Cultura biotecnológica en España: Análisis e interpretación de datos, Madrid, Genoma España.
WASSERMAN, S., y FAUST, K. (1994): Social Network Analysis, Cambridge, Cambridge University Press.
WITTEN, I. H. y E. FRANK (2005): Data Mining. Practical Machine Learning Tools and Techniques, Amsterdam, Elsevier.
WOOLDRIDGE, J.M. (2016): Introductory Econometrics: A Modern Approach, Australia, Cengage Learning.
ZHAO, Y. (2013): R and Data Mining: Examples and Case Studies. San Diego, Elsevier.