Proceso visual de clasificación de empresas mediante el uso de grafos y el censo industrial de Rafaela

Javier Fornari[1], Sergio Gramajo[2], Rodolfo Neira[3], Mariano Cordero[4], Melina Gaspoz[4], Agustín Cabaña[1]

[1]Universidad Tecnológica Nacional Facultad Regional Rafaela
Rafaela, Santa Fe – Argentina

[2]Universidad Tecnológica Nacional Facultad Regional Resistencia
Resistencia, Chaco – Argentina

[3]Universidad Tecnológica Nacional Facultad Regional San Francisco
San Francisco, Córdoba – Argentina

[4]Instituto Nacional de Tecnología Industrial
Rafaela, Santa Fe – Argentina

{javier.fornari, gonzalo.beltramino, agustin.cabana,}@frra.utn.edu.ar,
sergiogramajo@gmail.com, rneira@arnet.com.ar, {mcordero,
mgaspoz}@inti.gob.ar

Abstract. *This research aims at the analysis of data collected in the last industrial census of Rafaela city, in order to detect groups of companies and their interrelation. The data base of the answers to the questionnaire is preprocessed in Python to be later read in the Gephi software. The segmentation is done considering the modularity of the network and the evolution of the graph, as parameters that determine the effective links between the companies are altered.*

Resumen. *Esta investigación tiene como objetivo el análisis de los datos recopilados en el último Censo Industrial de la ciudad de Rafaela, con el fin de detectar grupos de empresas y su interrelación. La base de datos de las respuestas al cuestionario es preprocesada en Python para posteriormente ser leída en el software Gephi. La segmentación es realizada considerando la modularidad de la red y la evolución del grafo, conforme se alteran parámetros que determinan los enlaces efectivos entre las empresas.*

Keywords. *Base de Datos, Censo Industrial, Grafos, Modularidad.*

1. Introducción

La estructura económica actual de la ciudad de Rafaela incluye tanto una amplia producción industrial como una significativa producción agropecuaria, que de diversas maneras ha contribuido al desarrollo de actividades manufactureras. Actualmente, Rafaela constituye un polo de desarrollo industrial regional y es centro de la cuenta lechera más importante de Argentina y Sudamérica. La ciudad de Rafaela fue escenario de un vigoroso proceso de industrialización debido a la instalación y desarrollo en su planta urbana de industrias transformadoras de materias primas regionales. Puede decirse que el proceso industrial en Rafaela ha seguido, en líneas generales el ritmo de proceso de industrialización operado en la economía argentina. Así mismo, Rafaela muestra una estructura industrial con variados sectores que la convierten en un área productiva poli sectorial, por lo que no resultó tan afectada ante las diversas crisis económicas y sociales que se sucedieron en el país. No obstante, de la diversificación de actividades que conforman el tejido industrial local, los sectores de alimentos y bebidas y fabricación de autopartes y de maquinaria y equipos aglutinan el 69% del empleo total [Censo Industrial Rafaela 2006] [OCDE 1997] [Censo Industrial Rafaela 2012].

En este contexto cobran fuerza las teorías relacionadas con las ventajas asociadas al desarrollo de clusters y distritos industriales. Los clusters son concentraciones geográficas de empresas e instituciones interconectadas entre sí por un hilo conductor, el cual puede ser un determinado sector, y todos ligados por externalidades de diversos tipos [Porter 1999], mientras que los distritos industriales son entidades socioeconómicas fuertemente identificadas con un territorio geográfico y con una identidad cultural e históricamente determinada [Alburquerque 2015]. En líneas generales, estas teorías sostienen la idea de que, dentro de un determinado territorio, con actores vinculados básicamente por relaciones no sólo económicas, la convivencia de un grupo importante de empresas que logren interactuar de manera organizada y que puedan desarrollar la capacidad de asociarse, competir, cooperar, encadenarse, aprender y especializarse, explotando toda la cadena de valor de un determinado proceso productivo, constituye una estrategia que permitirá a las empresas involucradas superar los desafíos impuestos por el dinámico entorno, mediante el logro de altos niveles de eficiencia y competitividad.

Desde al año 2000, el Instituto de Capacitación y Estudios para el Desarrollo Local (ICEDEL) de Rafaela ha realizado un censo industrial de la ciudad reiterándolo cada 6 años. Actualmente posee el relevamiento de datos de 3 censos junto con sus informes y comparativos. Hoy en día están disponibles grandes volúmenes de datos de las empresas gracias a la amplia utilización de sistemas de información, que representan el backend de un número creciente de servicios y aplicaciones. En realidad, las organizaciones públicas y privadas reconocen el valor de los datos como un activo clave para entender profundamente fenómenos empresariales, económicos y sociales que sirven para mejorar la competitividad empresarial en un entorno dinámico [Batini et al 2006] [Fox et al 1994] [Madnick 2009]. En efecto, como señala [Fayyad et al 1996] cuando se introdujo el proceso KDD, "el valor de almacenar volúmenes de datos depende de nuestra capacidad para extraer información, eventos y tendencias útiles para soportar las decisiones y políticas basadas en el análisis estadístico y la inferencia." En los últimos

años, las técnicas de análisis de datos se han convertido en una parte esencial del proceso de descubrimiento de información ya que contribuyen para garantizar la credibilidad del proceso general de generación de conocimiento, haciendo que el razonamiento sobre los datos sea una preocupación muy importante [Fisher et al 2012] [Pasi et al 2013]. Asimismo, las comunidades industriales y académicos han dedicado un gran esfuerzo para hacer frente a los problemas de calidad de datos, su gestión y limpieza [Holzinger 2013].

Este trabajo surge en el marco de un proyecto de investigación realizado en la Universidad Tecnológica Nacional Facultad Regional Rafaela, que tiene por objetivo aplicar diferentes técnicas de minería de datos a la información obtenida en el último Censo Industrial Rafaela del año 2012, para obtener posteriormente una caracterización y clasificación de las diferentes empresas que componen la industria de la zona.

Las instituciones de la región, tales como la Municipalidad de la Ciudad de Rafaela, acompañan permanentemente a las empresas de la zona. Para llevar a cabo este trabajo es de interés contar con información sobre estas. Segmentando la totalidad de las empresas en pequeños grupos, con rasgos y necesidades similares, la tarea de generar planes de apoyo y consultoría resulta mucho más sencilla.

Los datos más recientes correspondientes al sector industrial de los que se dispone son del Censo Industrial Rafaela 2012, este último fue llevado a cabo por el ICEdeL (Instituto de Capacitación y Estudio para el Desarrollo Local), mientras en 2018 se está realizando el nuevo, pero aún no ha finalizado. El cuestionario utilizado por esta institución para las tareas de relevamiento está compuesto por dos formularios, con un total de 1078 preguntas, en el que participaron 497 industrias de la ciudad.

En un trabajo anterior aún no publicado, se realizó un análisis de clusters con métodos particionales y basados en la densidad, empleando variables numéricas con un bajo número de elementos perdidos, logrando una clasificación efectiva según los índices de validación empleados. En esa primera instancia se había prescindido de la utilización de variables categóricas para simplificar el análisis de los datos y principalmente, porque no poseían relevancia individualmente para realizar una segmentación.

El objetivo del trabajo presente es emplear 265 variables categóricas para establecer una división en grupos empleando el concepto de modularidad de una red, y a su vez, detectar que interrelación tienen las empresas de la industria local. La información del censo será previamente procesada con algoritmos en el lenguaje de programación Python con la ayuda de la librería Pandas para el manejo de datos la que facilita el tratamiento y limpieza de estos. Las redes tienen una representación gráfica que se denomina grafo, para lo cual se empleará un software gratuito y de fuente abierta denominado Gephi. Esta herramienta interactiva permite el estudio de los grafos en detalle sin escribir prácticamente ninguna línea de código y además con muchas opciones para el estudio y análisis de cualquier tipo de red.

2. Marco teórico

2.1. Grafos, nodos y enlaces

Se denomina grafo a la representación gráfica de una red. Un grafo consiste de dos elementos fundamentales, denominados vértices o nodos y lados o enlaces. Los enlaces

pueden representarse también mediante dos nodos, que son aquellos que se conectan entre sí [Guichard, 2018].

Un grafo es simple si cada uno de los nodos se unen por un único enlace, caso contrario, se denomina multigrafo. Además, los grafos pueden ser dirigidos o no dirigidos, teniendo en cuenta si los enlaces tienen un sentido asignado o no.

2.2. Modularidad

Normalmente, los nodos se ubican naturalmente en comunidades, altamente interconectadas interiormente y con unas pocas conexiones entre las diferentes agrupaciones. El problema radica en la separación de los nodos pertenecientes a un grupo, con respecto a los de otros [Newman 2006].

En el desarrollo del presente trabajo, la descomposición de la red en grupos se realiza utilizando un método de maximización de una función objetivo, la denominada modularidad. Esta se utiliza en sistemas complejos, para lograr dividirlos en partes que son más simplemente entendibles [Baldwin and Clark 2000]. La modularidad mide la densidad de los enlaces dentro de las comunidades, comparadas con los enlaces entre las diferentes comunidades. Si los enlaces tienen un peso, esta se define matemáticamente como se expresa a continuación.

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

En esta última ecuación A_{ij} define al peso del enlace que une los nodos i y j respectivamente. Por otro lado k_i y k_j representan la suma de los pesos de los enlaces vinculados a los nodos i y j. Además, c_i y c_j son las comunidades a la cual los nodos son asignados. La función $\delta(u, v)$ es igual a 1 si u y v son equivalentes, de otra forma es nula. Finalmente, m se define como la mitad de la sumatoria de todos los pesos de los enlaces que llegan a i y j respectivamente. La función modularidad Q toma valores que van desde -1 hasta 1, siendo considerado 0,3 en la práctica como un valor aceptable para la separación en comunidades [Blondel et al 2008].

3. Metodología

En los siguientes párrafos se explicará detalladamente el proceso realizado para llegar a los resultados. Esta serie de pasos se puede resumir en el diagrama de flujo que se expone a continuación en la Figura 1.

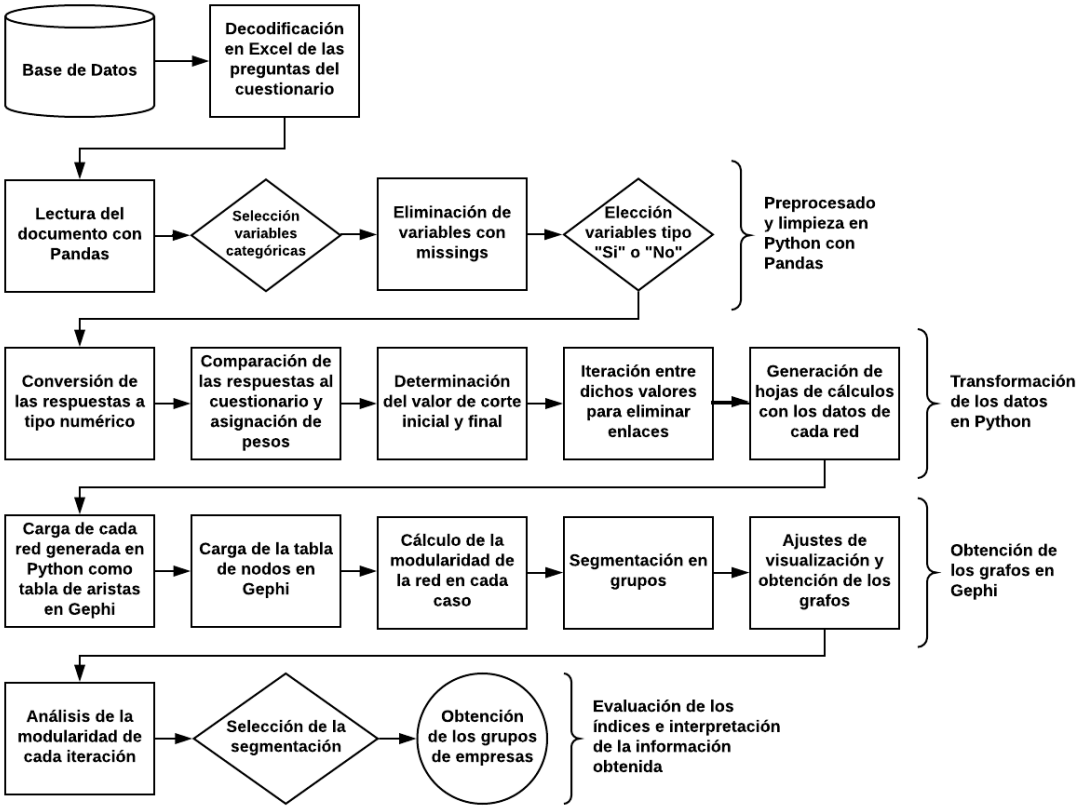

Figura 1. Resumen de la metodología empleada.

3.1. Preprocesamiento y limpieza de los datos

La base de datos del Censo Industrial Rafaela 2012 posee las respuestas al cuestionario de cada una de las 493 empresas participantes. Dicho cuestionario tiene 1078 preguntas de diversa índole, tales como económico, financiero, social, administrativo y tecnológico, entre otras.

Con respecto a las características de la base de datos, se puede mencionar que, debido a la naturaleza del cuestionario, no todas las preguntas eran de carácter obligatorio, por lo que muchas poseen un elevado número de elementos perdidos o missings. Por este motivo, y debido a que se desea realizar un proceso descriptivo y no predictivo, posteriormente se removerán algunas de ellas.

El proceso formal de obtención de información comienza con la identificación de las preguntas correspondiente al cuestionario, ya que las mismas, así como las respuestas se encuentran codificadas.

Luego, ocurre el procesamiento en Python de la base de datos. La hoja de cálculo con las preguntas, que de ahora en más consideraremos como variables es cargada utilizando la librería específica de Pandas.

Posteriormente, son seleccionadas todas las variables categóricas, que son aquellas cuyas respuestas no son numéricas. Del total, se filtran aquellas que no poseen

elementos perdidos, quedando 287 variables restantes. Se seleccionan 265 variables que son preguntas diversas al cuestionario cuyas respuestas corresponden por “Sí” o “No”.

3.2. Transformación de los datos

Las respuestas son transformadas a valores numéricos equivalentes a “1” para “Sí” y “2” para “No”. Se ejecuta un algoritmo que toma de a pares y sin repetir a las empresas, y compara los valores numéricos asignados a cada una de las variables. Se comparan las respuestas de una variable, y posteriormente si la diferencia entre las dos respuestas es nula, se aumenta en una unidad el peso asignado. Se continúa de esta manera con cada una de las variables, hasta completar las 265 comparaciones. Esto significa que el peso máximo que puede asignarse a un enlace corresponde a dicho valor, para el cual significa que el número de coincidencias en las respuestas es total. Por otro lado, si las empresas difieren totalmente en sus respuestas, el valor asignado al enlace es nulo. El proceso se repite, y mientras es ejecutado se almacena en una lista los enlaces cuyo peso ya ha sido calculado, de manera tal que solo se realicen de manera única los cálculos.

Así, se obtienen los pesos de cada uno de los enlaces entre dos empresas cualesquiera, que representan la similitud entre las respuestas al cuestionario evaluadas. En una primera instancia se obtienen 121278 enlaces con sus pesos correspondientes entre las 493 empresas. Este valor de conexiones surge del siguiente cálculo, correspondiente a la una matriz triangular de 493 filas por 493 columnas sin tener en cuenta la diagonal principal, dado por la ecuación expresada a continuación, en donde L es la cantidad de enlaces y N es la cantidad de nodos:

$$L = \frac{N(N+1)}{2} - N = \frac{493(493+1)}{2} - 493 = 121278$$

Este resultado parcial implica que la red se encuentra totalmente conectada, lo que significa que cada nodo tiene un enlace con los restantes. Para realizar una segmentación que la red este completamente interrelacionada es un limitante importante. Por este motivo, se establece la siguiente lógica para la eliminación de conexiones que permita una posterior segmentación en comunidades. Primero, se calcula el valor medio de todos los pesos de todos los enlaces, dando 200 como resultado redondeado al entero más cercano. En segundo lugar, se toma dicho valor como valor de corte inicial. Se denomina valor de corte a aquel, que si el peso del enlace existente entre dos empresas no alcanza dicho umbral, la similitud existente entre ellas representada por el peso no es suficiente para que el enlace sea considerado como efectivo, y por lo tanto es eliminado. En otras palabras, cualquier enlace de peso menor al valor de corte asignado es removido. Se genera una hoja de cálculo que posee la información sobre los nodos conectados, el peso del enlace, la etiqueta asignada al enlace y si el mismo es dirigido o no dirigido. Este archivo es el que luego se ingresará en Gephi como tabla de aristas para generar el grafo que representa a la red. En una tercera etapa se itera el segundo paso desde el valor inicial de corte hasta el valor máximo, es decir desde 200 hasta 265. Esto produce como salida redes con un número progresivamente menor de enlaces y nodos relacionados.

3.3. Cálculos de la modularidad y obtención de los grafos

Los archivos generados por la aplicación del algoritmo de Python son introducidos en Gephi. En este software se realiza el cálculo de la modularidad para cada una de las instancias y se producen los gráficos representativos de las redes. Se obtienen, además, el número de enlaces existentes y el número de nodos conectados luego de la eliminación. Cabe destacar que los grafos generados son simples, no dirigidos, y con enlaces con peso.

3.4. Generación de los grupos de empresas e interpretación de la información

Tomando en consideración la modularidad, se evalúa del total de iteraciones realizadas cual es la que posee los mejores índices. Considerando que el número de nodos involucrados y enlaces disminuye conforme aumenta el valor de corte, será aquella más próxima que sobrepase el valor de 0,3 de modularidad. En base a esto, se generan los grupos de empresas y se evalúan otras respuestas al cuestionario para obtener más información sobre la partición.

4. Resultados

En el siguiente apartado se exponen los resultados correspondientes a los índices para cada iteración del proceso, así como los gráficos representativos de las redes generadas.

4.1. Resultados de los índices para cada iteración

En la Tabla 1 se expresan los valores de los índices de modularidad (Q), número de nodos (N), fracción del total de nodos vinculados por al menos un enlace (N%), número de enlaces (L) y fracción del total de enlaces con respecto a una red totalmente conectada (L%). Cada uno de ellos obtenidos para las iteraciones del valor de corte (VC). Nótese que el paso de VC es unitario debido a que los pesos también son de tipo enteros. N% se obtiene al dividir el número de nodos para un determinado VC que se encuentran interconectados sobre el total de nodos interconectados posibles, que es 493. L%, por su parte, es el cociente entre el total de enlaces dados para un determinado VC y el número máximo de enlaces para una red totalmente conectada, es decir 121278 enlaces.

Tabla 1. Índices correspondientes a la evolución de la red

VC	Q	L	L%	N	N%
200	0,033	67457	0,556217946	474	0,961
…	…	…	…	…	…
210	0,037	48586	0,400616765	433	0,878
…	…	…	…	…	…
220	0,050	29749	0,245295932	371	0,753
…	…	…	…	…	…
230	0,068	14515	0,119683702	285	0,578
…	…	…	…	…	…
240	0,099	4438	0,036593611	192	0,389
241	0,104	3854	0,031778229	172	0,349
242	0,103	3281	0,027053546	167	0,339
243	0,100	2756	0,022724649	152	0,308
244	0,111	2312	0,019063639	142	0,288

245	0,123	1925	0,015872623	132	0,268
246	0,120	1613	0,013300021	120	0,243
247	0,137	1327	0,010941803	116	0,235
248	0,143	1080	0,00890516	106	0,215
249	0,154	834	0,006876762	97	0,197
250	0,149	646	0,005326605	84	0,170
251	0,173	491	0,00404855	75	0,152
252	0,191	363	0,002993123	66	0,134
253	0,238	263	0,002168571	58	0,118
254	0,288	182	0,001500684	49	0,099
255	0,329	123	0,001014199	42	0,085
256	0,361	89	0,000733851	38	0,077
257	0,371	55	0,000453504	32	0,065
258	0,514	36	0,000296839	26	0,053
259	0,606	21	0,000173156	21	0,043
260	0,695	11	9,07007E-05	14	0,028
261	0,693	7	5,77186E-05	10	0,020
262	0,560	5	4,12276E-05	7	0,014
263	0,375	4	3,29821E-05	5	0,010
264	0,500	2	1,6491E-05	4	0,008
265	0	1	8,24552E-06	2	0,004

Tal como se observa en la Tabla 1, a medida aumenta el valor de corte VC, el número de enlaces L que inicialmente es de 67457 disminuye progresivamente hasta 2 debido a la eliminación de los enlaces que son menores a VC. De igual manera, el número de nodos N que tienen al menos una conexión con otro nodo cae conforme aumenta la eliminación de enlaces. Sin embargo, la modularidad Q tiene un comportamiento inverso, aumentando en paralelo con VC, hasta un pico de 0,695 para un VC equivalente a 260. Las modularidades más altas se dan entre los VC 258 y 262. Sin embargo, estos puntos individuales del proceso no son óptimos, ya que la eliminación de nodos es demasiado alta.

Dicho comportamiento se visibiliza más fácilmente en la Figura 2, expuesta a continuación. Debido a la escala, se exponen los porcentajes de $L_{\%}$ y $N_{\%}$ como fracción entre cero y la unidad.

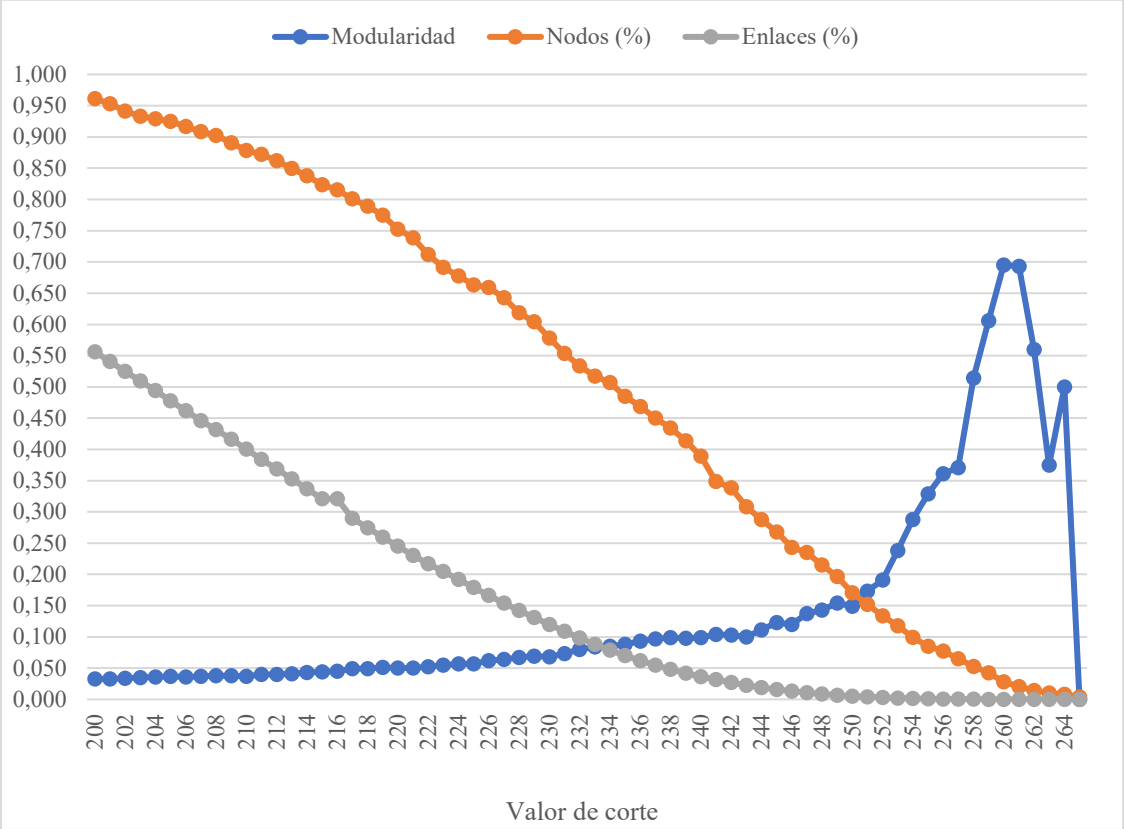

Figura 2. Q, $N_{\%}$ y $L_{\%}$ conforme aumenta VC.

4.2. Grafos generados

A continuación, se muestran en las Figuras 3 a 9 los gráficos de las redes de empresas generadas. Debido a que el intervalo a describir es amplio, solo se exponen las agrupaciones generadas por modularidad en los valores de corte múltiplos de 5.

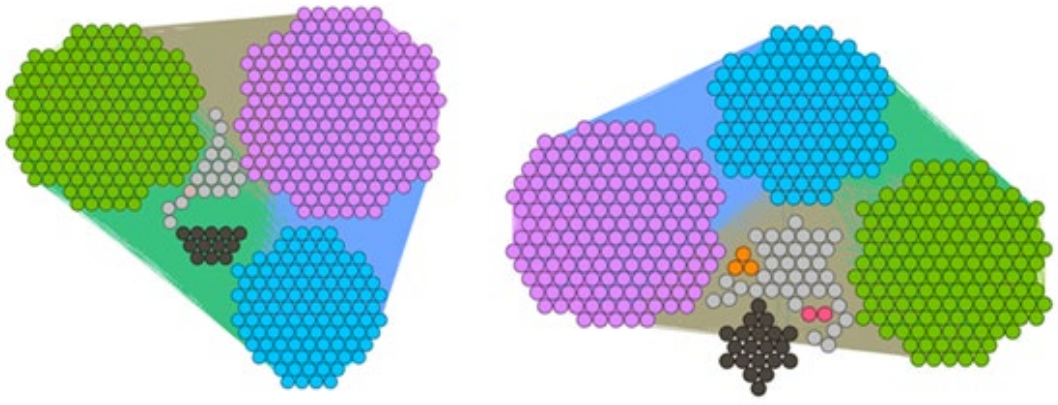

Figura 3. Grafos para valor de corte igual a 200 (izquierda) y 205 (derecha).

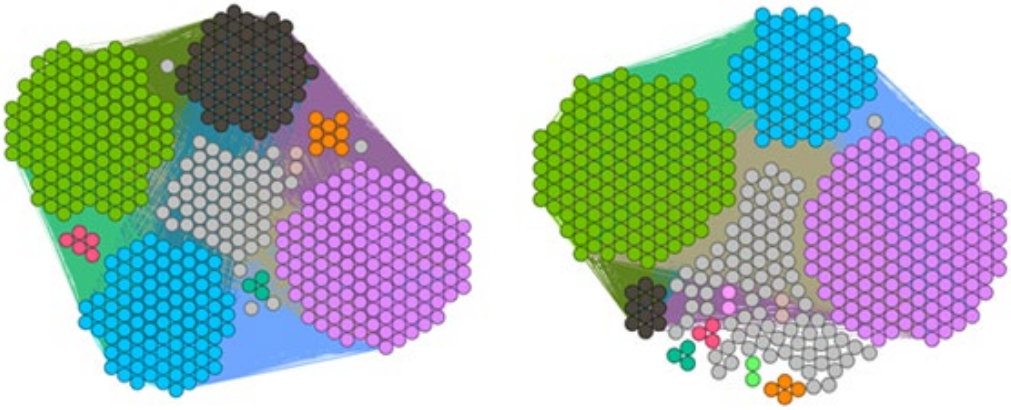

Figura 4. Grafos para valor de corte igual a 210 (izquierda) y 215 (derecha).

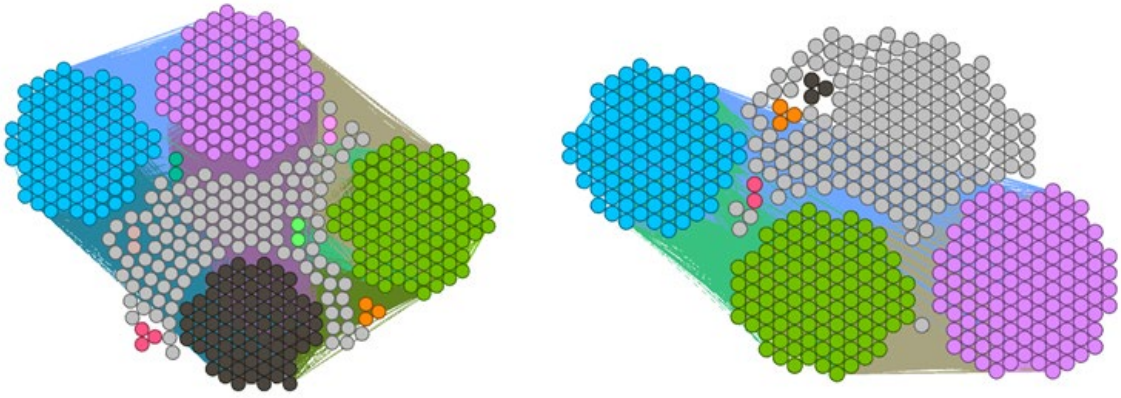

Figura 5. Grafos para valor de corte igual a 220 (izquierda) y 225 (derecha).

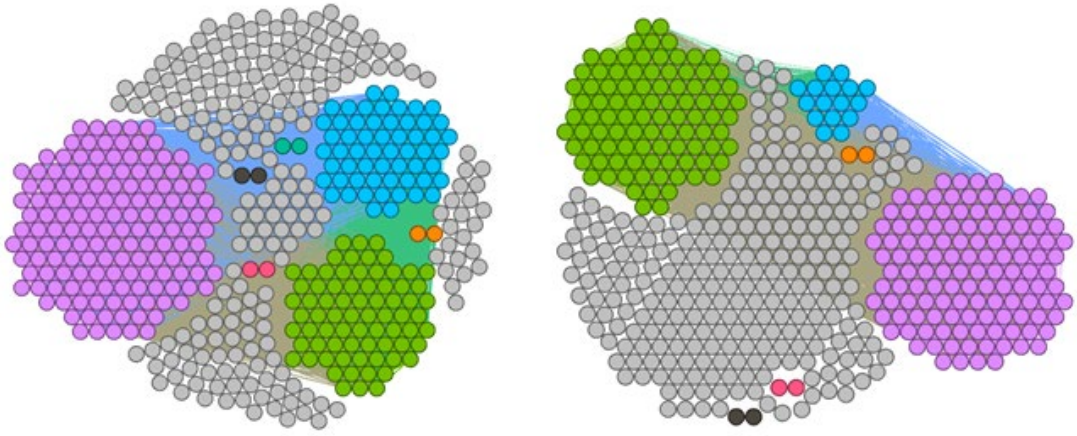

Figura 6. Grafos para valor de corte igual a 230 (izquierda) y 235 (derecha).

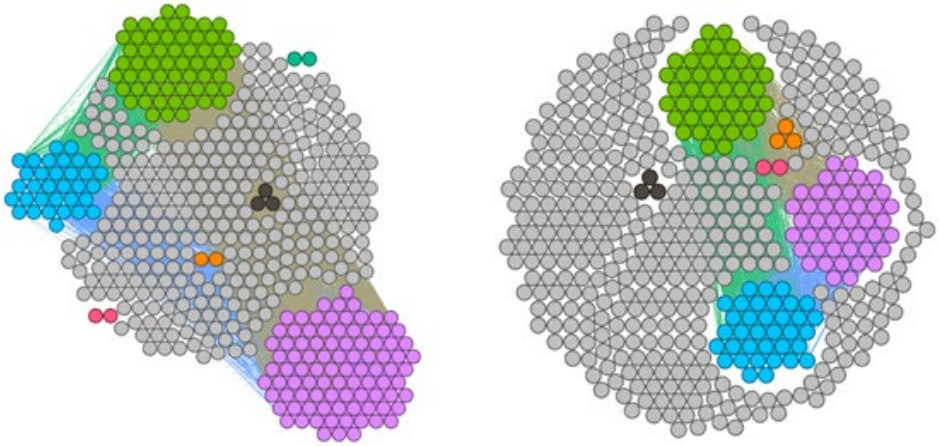

Figura 7. Grafos para valor de corte igual a 240 (izquierda) y 245 (derecha).

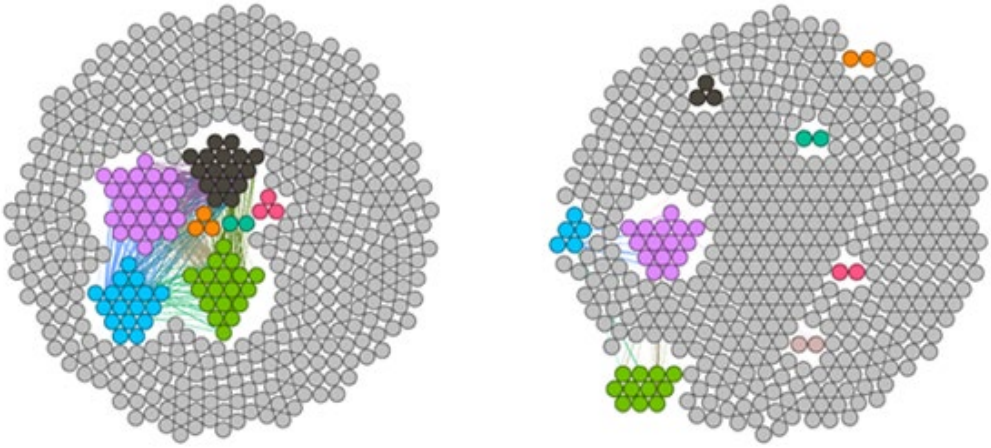

Figura 8. Grafos para valor de corte igual a 250 (izquierda) y 255 (derecha).

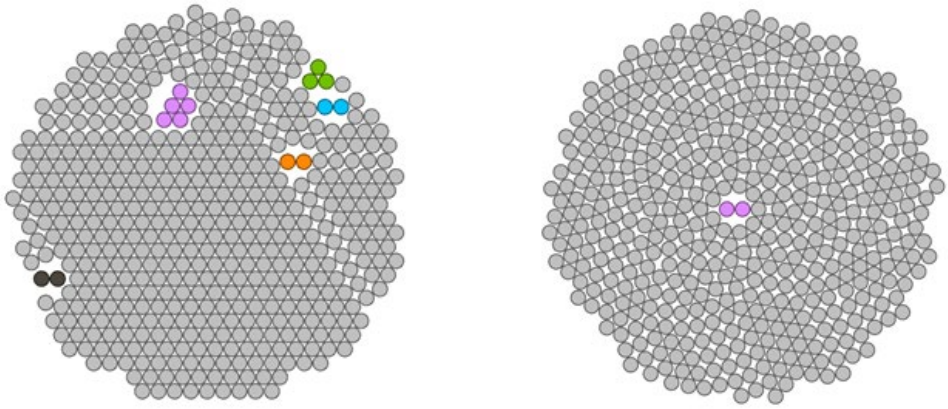

Figura 9. Grafos para valor de corte igual a 260 (izquierda) y 265 (derecha).

Cabe destacar que los colores asignados no son representativos a ningún grupo en particular, sino que solo se utilizan con el objetivo de diferenciarlos. Los nodos que no se encuentran conectados, se representan con el color gris.

En las figuras anteriores se observa que a medida que crece el valor de corte, disminuyen los enlaces, y en consecuencia los grupos se vuelven más reducidos. A su vez, se generan más nodos aislados.

A continuación, se procederá a realizar un análisis descriptivo de dichas imágenes que expresan la evolución de la red.

Al observar la figura 3, para la iteración correspondiente al valor de corte igual a 200, se observan tres grupos principales y uno pequeño, con poca presencia de nodos aislados en gris. Para un valor de corte de 205 la situación es similar, con la particularidad de la aparición de dos grupos minoritarios adicionales.

En el caso de la figura 4, para el valor de corte de 210, existen cuatro grupos mayoritarios y cinco menores, mientras que para el valor de corte de 215 desaparece un grupo mayoritario y se incrementa en una unidad el número de grupos minoritarios. En ambas iteraciones se añaden nodos aislados.

Por su parte, en la figura 5 se repiten situaciones similares a las vistas en la figura 4, con la peculiaridad de que el número de grupos pequeños es menor.

Para la figura 6 se puede destacar como en ambas iteraciones el número de nodos aislados comienza a superar el número de nodos que se encuentran interconectados en grupos, situación que es confirmada numéricamente por la información aportada en el gráfico de líneas de la figura 2.

En la figura 7 se puede visualizar en ambos casos tres grupos mayoritarios y la presencia de cuatro grupos minoritarios para el valor de corte de 240 y tres grupos para el valor de corte de 245.

Respecto a la figura 8, para el valor de corte de 250 se observa una clara disminución del tamaño de los grupos generados, mientras que las agrupaciones dadas para el valor de corte equivalente a 255 son constituidas por pocos nodos, siendo solo dos de ellas un tanto mayor.

Finalmente, en la figura 9 solo se pueden observar, tanto para la iteración con valor de corte de 260 como la de 265, grupos muy pequeños de nodos, con pocas conexiones, pero cuyos pesos indican que están fuertemente relacionados. En la iteración de 265 se observa solo un grupo de dos empresas, que coincidieron en absolutamente todas las respuestas al cuestionario y, por lo tanto, el enlace posee el peso máximo asignable.

4.3. Segmentación en grupos de empresas

Si bien para valores de corte cercanos al extremo inferior se producen agrupaciones que involucran a prácticamente la totalidad de las empresas, el valor obtenido en la modularidad es demasiado cercano a cero para ser consideradas como certeras. Por este motivo, se considera trabajar con valores de modularidad inmediatos y superiores a 0,3. Dicho valor corresponde a la iteración con valor de corte igual a 255. Se demostrará que, para valores de corte menores, la segmentación no produce grupos estables ni con características definidas.

El grafo correspondiente al valor de corte 200 tiene una modularidad igual a 0,033. Al ser cercano a cero, esto significa que la distribución de los enlaces y nodos no permite diferenciar efectivamente grupos, por lo que la partición obtenida es prácticamente

aleatoria. Este conjunto de grupos tiene para el primer cálculo de modularidad la siguiente representación gráfica de la Figura 10.

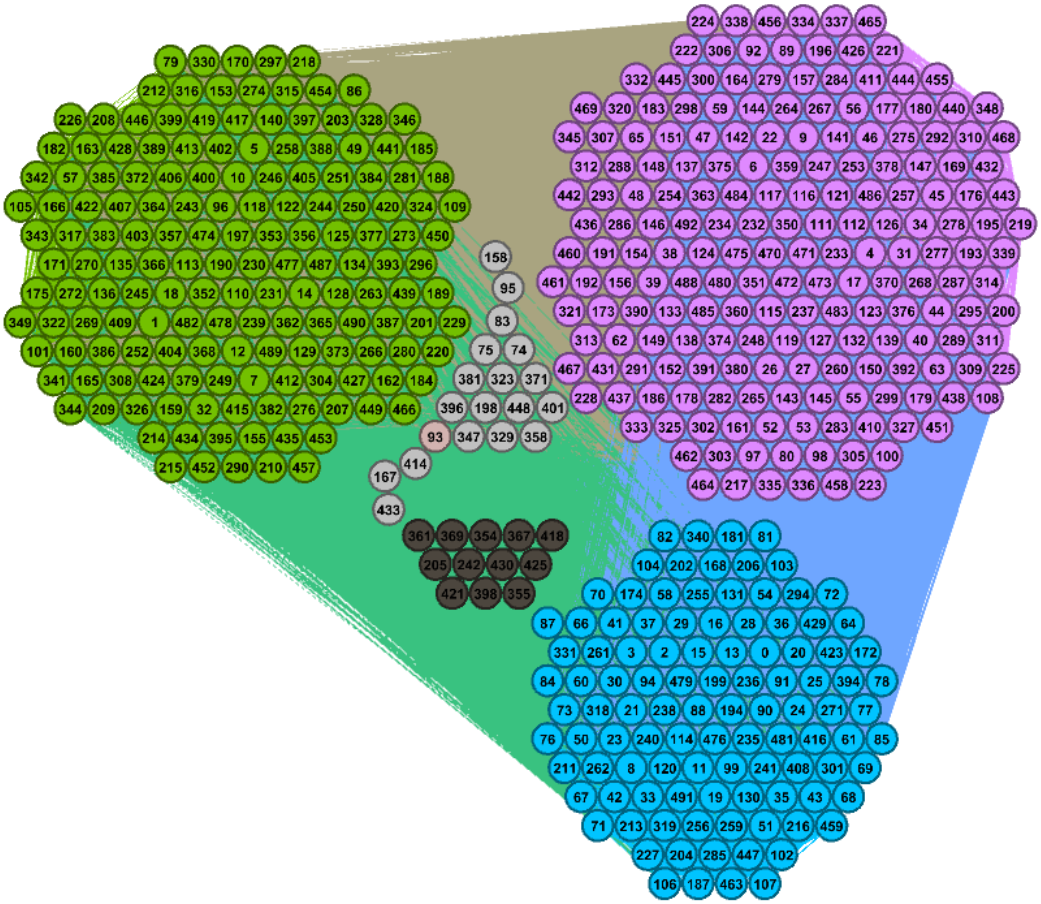

Figura 10. Grafos para valor de corte igual a 200, primera distribución generada.

En una primera instancia, se generan los grupos de acuerdo a la modularidad, siendo tanto el color como la agrupación concordantes con esta segmentación. Sin embargo, si mantenemos los nodos en sus respectivos lugares y se recalcula la modularidad, se generan nuevos conjuntos. Como vemos en la Figura 11, si bien ciertas empresas siguen perteneciendo al mismo grupo, otras varían.

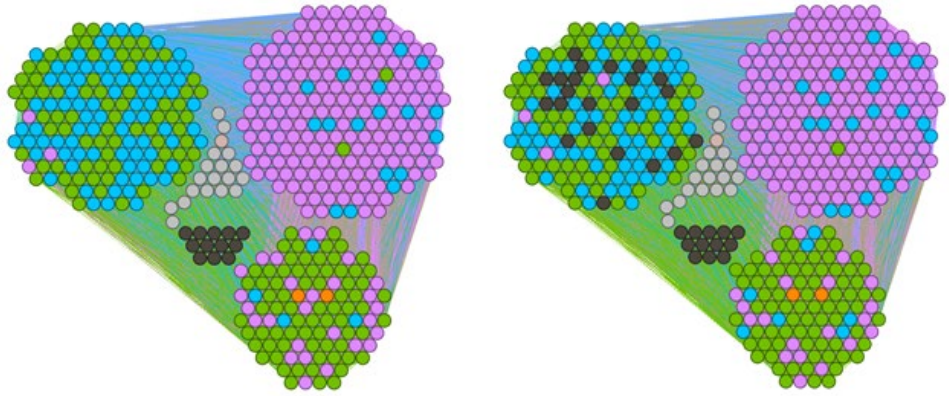

Figura 11. Grafos para valor de corte igual a 200, recalculando la modularidad.

Esto confirma los resultados de la modularidad, y que dicho valor al ser tan bajo no permite establecer grupos claramente diferenciados.

Si se efectúa el mismo proceso para la red con valor de corte igual a 255, cuya modularidad es 0,329 se obtiene para un primer cálculo de la modularidad la Figura 12.

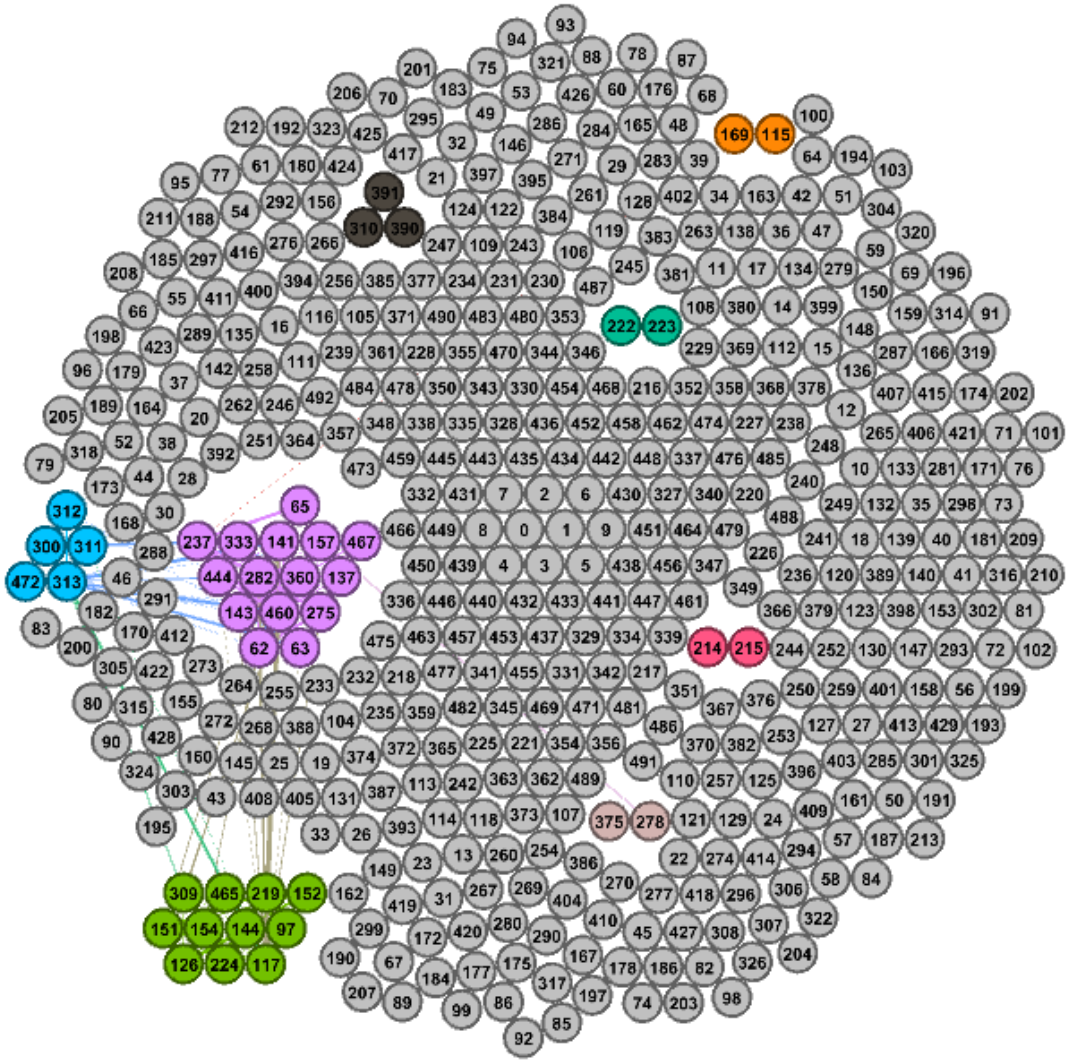

Figura 12. Grafos para valor de corte igual a 255, primera distribución generada.

En cálculos de modularidad posteriores, las agrupaciones permanecen prácticamente constantes, sin demasiados pasajes de nodos entre los grupos, tal como se observa en la Figura 13.

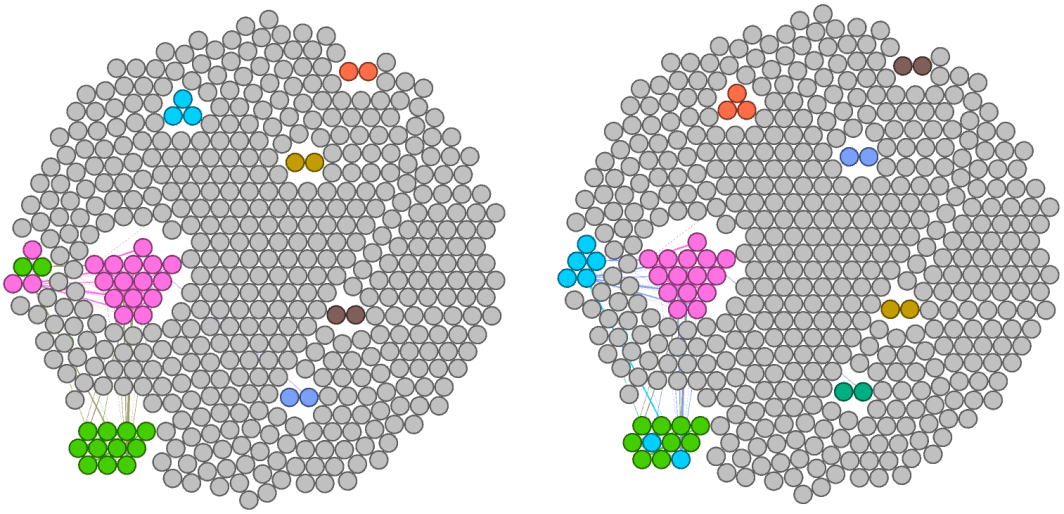

Figura 13. Grafos para valor de corte igual a 255, recalculando la modularidad.

Tal como se puede ver en la Tabla 1, en este segundo caso la modularidad es mucho más alta y, por ende, la segmentación más estable a costa de la eliminación de nodos del análisis. Se tomará como modelo para la clasificación esta última iteración, representada por la Figura 12.

4.4. Caracterización de las agrupaciones en base a información externa

A continuación, se empleará información que no fue suministrada al algoritmo generado en Python cuya salida es introducida en Gephi para la obtención de los grafos. De esta manera, se puede realizar una validación cruzada del proceso y evaluar si efectivamente los grupos comparten características en común, además de las coincidencias que son evidenciadas por el peso de los enlaces.

La información que se utilizará en este proceso consiste en algunas de las variables categóricas descartadas inicialmente para la formación de las redes y adicionalmente variables de tipo numéricas empleadas en un anterior trabajo de análisis de clusters con esta base de datos. Se utilizan estas variables por su relevancia y además, porque no poseen missings. Si bien existen otras variables que pueden emplearse, no son significativas a la hora de caracterizar aspectos de una empresa.

De manera concreta, se expresan en la Tabla 2 el nombre y tipo de variables que serán utilizadas. Además, se le asigna una letra identificatoria a cada variable que será posteriormente empleada.

Tabla 2. Variables empleadas para la caracterización de los grupos.

ID	Nombre de la variable	Tipo de variable
A	Primer producto de la empresa	Categórica
B	Primera debilidad de la empresa	
C	Primera amenaza de la empresa	
D	Primera oportunidad de la empresa	
E	Personal ocupado en el año 2012	Numérica
F	Endeudamiento respecto a la facturación en el año 2012	

Teniendo en cuenta esto y la segmentación lograda que se expone en la Figura 12, los grupos formados y las características a evaluar seleccionadas se muestran a continuación en las Tablas 3 a la 9. El número de empresa (NE), es el correspondiente a la figura 12.

Tabla 3. Características grupo 1.

NE	A	B	C	D	E	F
62	Pan	Falta de personal capacitado	Muchos competidores que tienen mejor captación de clientes	No supo identificar oportunidades	2	50
63	Churros	Falta solidez financiera	No supo identificar amenazas	No supo identificar oportunidades	1	0
65	Pan	Personal irresponsable	Bajo precio de mercado del producto	No supo identificar oportunidades	5	5
137	Remeras	Cuenta corriente con clientes	No supo identificar amenazas	No supo identificar oportunidades	1	0
141	Producción vainas cuchillos	Bajo volumen de producción	Reemplazo de artículos de cuero artesanales por industriales	No supo identificar oportunidades	1	0
143	Producto intermedio cepillado de madera	No se auto provee la materia prima y debe comprarla.	Poca oferta de materia prima	Considera que no tiene	2	0
157	Techos de madera	Falta maquinaria moderna	Dificultad para conseguir materia prima	No supo identificar oportunidades	2	5
237	Tapiales prefabricados de hormigón	Falta equipo de trabajo	Alto costo de materia prima	Construcción planes de vivienda	4	0
275	Rejas	No supo identificar debilidades	No supo identificar amenazas	No supo identificar oportunidades	1	0
282	Aberturas de todo tipo en hierro (puertas/	Maquinaria obsoleta	Ausencia de financiamiento	No supo identificar oportunidades	1	0

	ventanas/ portones)					
333	Canaletas	Trabajo manual (falta tecnología)	Falta de financiamiento	No supo identificar oportunidades	1	0
360	Bombas de freno mecanizadas	Dificultad en cobrar	Gran cantidad de competencia	No supo identificar oportunidades	2	3
444	Placares en madera	No supo identificar debilidades	Inflación	Reparación de maquinarias	2	0
460	Muebles de cocina de madera	Falta maquinaria modernizada	Gran cantidad de competencia	No supo identificar oportunidades	1	5
467	Redes de futbol	Trabajo manual (falta tecnología)	No supo identificar amenazas	No supo identificar oportunidades	1	0

Como se puede ver en la Tabla 3, el grupo 1 cuenta con una gran cantidad de empresas dedicadas a materiales y productos de construcción (6) y otros que se dedican a la fabricación de muebles de madera (2). También hay un conjunto de empresas que se dedican a la panadería (3). Las empresas restantes (4) pertenecen a distintos rubros. Con respecto a la cantidad de empleados, todas oscilan entre unipersonales y cinco empleados, así como también mantienen un nivel bajo de endeudamiento en su mayoría. Con respecto a las oportunidades, la mayoría de ellas no pudo identificarlas, mientras que en las debilidades varias empresas (6) expresaron dificultades en conseguir maquinaria y tecnología necesaria. Las amenazas se encuentran divididas entre la dificultad de conseguir materia prima, la competencia y la falta de financiamiento.

Tabla 4. Características grupo 2.

NE	A	B	C	D	E	F
97	Soda	No supo identificar debilidades	No supo identificar amenazas	No supo identificar oportunidades	1	0
117	Trapos de piso	Poca participación en el mercado	Competencia excesiva	Posibilidades de financiamiento	1	0
126	Remeras	No supo identificar debilidades	No supo identificar amenazas	No supo identificar oportunidades	1	0
144	Maderas aserradas de cedro	Falta de iniciativa	Competidores	Actualización tecnológica	1	0

151	Aberturas en maderas para casas (puertas, portones, etc)	Maquinaria antigua	No supo identificar amenazas	Espera ser proveedor de municipalidad	1	20
152	Hojas puertas placas de madera	Maquinaria antigua	No supo identificar amenazas	No supo identificar oportunidades	1	0
154	Puertas placas	Falta de luz trifásica	Inflación	No supo identificar oportunidades	1	20
219	Ladrillos comunes	No supo identificar debilidades	Inflación	Ubicación física	1	0
224	Lajas premoldeadas	Fallecimiento del personal	No supo identificar amenazas	No supo identificar oportunidades	1	0
309	Servicio de trabajo de metales	No supo identificar debilidades	No supo identificar amenazas	No supo identificar oportunidades	1	0
465	Alianzas de oro y plata	Acceso a la infromática (interna)	Incertidumbre en cuanto al mercado	No supo identificar oportunidades	1	0

La situación del grupo 2 expresada por la Tabla 4 es similar a la anterior en cuanto a la identificación de las oportunidades, la mayoría de las empresas no pudo efectuarlas. En cuanto a los rubros, se destaca un grupo de empresas dedicadas a materiales y elementos de construcción (6), mientras que las demás empresas son pertenecientes a otros sectores. Las debilidades principales consisten en problemas con la maquinaria y las instalaciones (4), mientras que las amenazas no pudieron ser identificadas por muchas empresas (6) y las restantes se dividen entre competencia (2), inflación (2) y otras (1).

Tabla 5. Características grupo 3.

NE	A	B	C	D	E	F
300	Mantenimiento (tornería general)	No supo identificar debilidades	No supo identificar amenazas	No supo identificar oportunidades	1	0
311	Servicio metalúrgico para empresas (tornería)	Alto endeudamiento	Falta de créditos para su empresa	Crecimiento del mercado	1	0
312	Reparaciones en general de tornería (brinda	No supo identificar debilidades	Falta de trabajo	No supo identificar oportunidades	1	0

	servicio de reparación)					
313	Productos mecanizados	Problemas financieros	Incertidumbre de continuar en el mercado por juicio muy caro	No supo identificar oportunidades	1	0
472	Escobas	No supo identificar debilidades	Competencia	No supo identificar oportunidades	1	0

El grupo 3 presenta claras características similares, la mayoría de las empresas es del rubro de la metalmecánica (4) mientras que la otra empresa restante comercializa escobas. Todas estas empresas son unipersonales y el nivel de endeudamiento para el año 2012 era nulo. Con respecto a las oportunidades, debilidades y amenazas son variadas.

Tabla 6. Características grupo 4.

NE	A	B	C	D	E	F
310	Repuestos para plantas de alimento balanceado	No supo identificar debilidades	Gobierno nacional	No supo identificar oportunidades	1	0
390	Cortadora de cesped	Falta de solidez financiera	Falta de acceso al financiamiento	Experiencia en el mercado	42	25
391	Agropartes (repuestos para máquinas agrícolas)	Altos costos fijos	Inflación	Oportunidades de desarrollo comercial a partir de las ventas	2	20

El grupo 4 solo presenta una leve coincidencia entre los rubros que caracterizan a las empresas. En el resto de los ítems no se observa ninguna concordancia.

Tabla 7. Características grupo 5.

NE	A	B	C	D	E	F
222	Macetas de cerámicas	Alta variación en las ventas	Inestabilidad económica nacional	No supo identificar oportunidades	1	10
223	Macetas de cerámica	Alta variación en las ventas	Inestabilidad económica nacional	No supo identificar oportunidades	1	10

En el caso de las dos únicas empresas correspondientes al grupo 5, la coincidencia entre las distintas variables es total, tal como se puede apreciar. Se dedican al mismo rubro, poseen las mismas debilidades y amenazas, y son empresas individuales.

Tabla 8. Características grupo 6.

NE	A	B	C	D	E	F
278	Puertas de hierro y chapa	Problemas con organización de los tiempos de producción	Inflación	No supo identificar oportunidades	2	5
375	Transmisión de velocímetro	Dificultad para conseguir personal capacitado	Gran cantidad de competencia	No supo identificar oportunidades	2	3

En esta última tabla, que representa las características de las empresas del grupo 6, no se puede dilucidar ningún patrón coincidente entre sus integrantes entre las preguntas que integran las variables categóricas. Sin embargo, ambas tienen un nivel bajo de endeudamiento y están compuestas por dos personas.

Tabla 9. Características grupo 7.

NE	A	B	C	D	E	F
214	Inyección de piezas	Comercialización	Falta de créditos bancarios	Ubicación física	14	10
215	Inyección de piezas	Comercialización	Falta de créditos bancarios	Ubicación física	5	2

Al igual que en el caso del grupo 5, el grupo 7 posee una coincidencia total entre las preguntas A, B, C y D, y difiere en el endeudamiento y el número de empleados.

5. Conclusiones

Se puede establecer que, al observar los casos de las empresas que componen los grupos 5 y 7, la gran cantidad de coincidencias entre variables externas a la segmentación por modularidad da un indicio acerca de que empresas de características similares responden de manera similar al cuestionario. Esto se visualiza en menor medida en el caso de los grupos 1 y 3, con coincidencias parciales. Por otro lado, utilizando las variables A, B, C, D, E y F que se plantean, las similitudes en las características de las empresas 4 y 6 son prácticamente nulas. Sin embargo, solo se ha empleado un subgrupo de variables categóricas que se consideran de relevancia, en un futuro trabajo podrían emplearse las restantes para evaluar este punto en mayor profundidad.

Con respecto a la segmentación lograda, el valor de la modularidad es suficiente para respaldar los resultados obtenidos, y la variación de nodos que constituyen los grupos es insignificante. Otro punto por destacar es que al aumentar la modularidad, se sacrifican

nodos que entrarían en el análisis de los grupos, pero se aumenta la estabilidad de las agrupaciones.

Como posibles aplicaciones de la información recopilada, se puede mencionar la utilización de esta segmentación para la orientación de políticas por parte del estado o de instituciones locales. Al considerar que los grupos se definen con enlaces cuyos pesos miden el nivel de coincidencia, los conjuntos de empresas logrados respondieron de manera muy similar al Censo Industrial y por lo tanto tienen características y problemas comunes, que pueden definirse en mayor detalle al evaluar cada pregunta del cuestionario. Más particularmente, por ejemplo, en el grupo 7 se presenta como amenaza la falta de créditos bancarios. En las empresas del grupo 1, se detecta un claro problema como debilidad, que es la falta de maquinaria y tecnología. Es ahí, donde el estado trabajando en conjunto con diversas organizaciones debe enfocar sus esfuerzos.

También sirve para la detección de empresas muy parecidas, a las que se les podría plantear debido a que están en rubros similares, acuerdos de cooperación o de trabajo en conjunto en ciertas áreas, como el desarrollo de tecnología.

6. Referencias

Alburquerque F., Cluster, Territorio y desarrollo empresarial: diferentes modelos de organización productiva, BID/FOMIN, http://www.iadb.org/mif/v2/fourthworkshoppipCR.htm , visto por última vez Mayo 2015

Baldwin, C. & Clark, K. (2000). "Design Rules: The power of modularity", páginas 63–92, MIT Press.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM Comput. Surv., 41, 16:1–16:52.

Blondel, V., Guillaume, J., Lambiotte, R. & Lefebvre, E. (2008). "Fast unfolding of communities in large networks", Journal of Statistical Mechanics: Theory and Experiment.

Censo Industrial Rafaela (2006), visto por última vez Febrero 2015, http://www.rafaela.gov.ar/nuevo/Files/Archivos/arc_93.pdf

Censo Industrial Rafaela (2012), visto por última vez Marzo 2015, https://www.rafaela.gov.ar/File.aspx?n=jQ7YxAsBkg2jTYtov**dNqVO6XE22AMhj6WVQQRUmndRloTvVbk6Mm4o3PjquvQf9&t=Informe_Final_Censo_Industrial_de_Rafaela_2012

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data. Communications of the ACM, 39, 27–34.

Fisher, C., Lauría, E., Chengalur-Smith, S., & Wang, R. (2012). Introduction to information quality. AuthorHouse. [17] Pasi, G., Bordogna, G., & Jain, L. C. (2013b). Quality issues in the management of web information. Intelligent systems reference library (Vol. 50). Springer.

Fox, C., Levitin, A., & Redman, T. (1994). The notion of data and its quality dimensions. Information Processing & Management, 30, 9–19.

Guichard, D. (2018). "An Introduction to Combinatorics and Graph Theory", página 91, Whitman College.

Holzinger, A., Yildirim, P., Geier, M., & Simonic, K.-M. (2013). Quality-based knowledge discovery from medical text on the web. In Pasi, Bordogna, and Jain.

Madnick, S. E., Wang, R. Y., Lee, Y. W., & Zhu, H. (2009). Overview and framework for data and information quality research. Journal of Data and Information Quality, 1, 2:1–2:22.

Newman, M. (2006). “Modularity and community structure in networks”, páginas 8577-8582, vol. 103, PNAS.

Organization for Economic Cooperation and Development (OCDE) (1997), Proposed Guidelines for Collecting and Interpreting Technological Innovation Data, , Paris, segunda edición.

Pasi, G., Bordogna, G., & Jain, L. C. (2013). Quality issues in the management of web information. Intelligent systems reference library (Vol. 50). Springer.

Porter M., (1999), Los clusters y la competitividad, en Elgue, M.C. (Eds.), Globalización, desarrollo local y redes asociativas, Edit. Corregidor.