REDES- Revista hispana para el análisis de redes sociales
Vol.6, #5, Jun./Jul 2004
http://revista-redes.rediris.es

Estimación y evaluación de Modelos estructurales centro-periferia

García Muñiz, Ana Salomé; Ramos Carvajal, Carmen; Álvarez Herrero, Rubén;

Fernández Vázquez, Esteban – Universidad de Oviedo[1]

Resumen

Resulta revelador intentar plasmar la idea referida a un conglomerado de agentes que constituyan un núcleo alrededor del cual gire la actividad objeto de estudio. La concepción de una estructura formada por un centro y una periferia, constituye un paradigma clásico y recurrente en muchos campos de la ciencia.

Siguiendo esta línea, los investigadores Stephen Borgatti y Martin Everett desarrollan un modelo estructural en 1999 basado en la delimitación de un centro formado por un conjunto de actores fuertemente relacionados, esto es, un grupo cohesivo y con alta densidad de interrelaciones. En contraposición, los agentes dispersos y poco conectados de la red delimitan la periferia del sistema.

El enfoque original de los autores es modificado empleando medidas que creemos, aportan un mayor grado de coherencia y exactitud a los objetivos planteados. En este trabajo, sin perdida de generalidad nos centramos en la estimación y posteriormente en la evaluación, de modelos centro-periferia basados en grafos valorados.

Palabras clave: redes sociales, análisis estructural, modelos centro-periferia.

Abstract

The conception of a structure made up of a core and periphery is a common, classic paradigm in many fields of science. Following this line, in 1999 researchers Stephen Borgatti and Martin Everett developed a model of structural analysis based on the delimitation of a core formed by a group of densely connected actors, in contrast to a class of actors, more loosely connected and forming the periphery of the system.

The original approach of these authors is modified, employing measures that, in our opinion, show a larger degree of coherence and accuracy in the proposed objectives.

Key words: network theory, structural analysis, core-periphery models.

[1] Enviar correspondencia a: Ana Salomé García Muñiz asgarcia@uniovi.es

1. Introducción

Resulta revelador intentar plasmar la idea referida a un conglomerado de agentes que constituyan un núcleo alrededor del cual gire la actividad objeto de estudio. La concepción de una estructura formada por un centro y una periferia, constituye un paradigma clásico y recurrente en muchos campos de la ciencia.

Siguiendo esta línea, los investigadores Stephen Borgatti y Martin Everett desarrollan en 1999 un modelo de análisis estructural basado en la delimitación de un centro formado por un conjunto de actores fuertemente relacionados, esto es, un grupo cohesivo y con alta densidad de interrelaciones. En contraposición, los agentes dispersos y poco conectados de la red delimitan la periferia del sistema.

El enfoque original de los autores es modificado empleando medidas que creemos, aportan un mayor grado de coherencia y exactitud a los objetivos planteados. En este trabajo, sin perdida de generalidad[2] nos centramos en la estimación y posteriormente en la evaluación, de modelos centro- periferia basados en grafos valorados. Se concluye con una breve síntesis de los resultados obtenidos.

2. Modelo centro-periferia. Planteamiento general

Siguiendo la línea de desarrollo iniciada por Borgatti y Everett (1999), es posible reconstruir un modelo donde se determine el centro o núcleo de la actividad frente a aquellos nodos que constituyen la periferia de la misma. La delimitación de las zonas de interés se basa en la comparación de una estructura ideal con los datos disponibles de la red, asumiendo que aquellos actores con unas relaciones más intensas serán los que constituyan el núcleo central.

La medida de proximidad entre la estructura real y teórica propuesta por Borgatti y Everett es:

$$\rho = \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} \delta_{ij}$$

[2] Las técnicas de fiabilidad empleadas resultan extensibles a grafos dicotómicos.

$$\delta_{ij} = c_i c_j$$

Donde x_{ij} muestra la presencia de relaciones observadas entre los nodos i-*ésimo* y j-*ésimo*, δ_{ij} indica la existencia de interrelaciones entre los actores en la matriz imagen ideal centro-periferia, y c_i es el grado de centralidad del actor i-*ésimo*, tal que $c_i \geq 0$.

La estructura teórica contendrá valores relativamente elevados para aquellos pares de actores que mantengan un alto nivel de centralidad (*coreness*), mientras que valores intermedios corresponderán a aquellos casos en los que sólo uno de los nodos muestra una posición central, por último, los actores pertenecientes a la periferia llevarían aparejados valores de pequeña cuantía.

La estimación de los valores correspondientes al centro (c_i), se efectúa a partir de un proceso de maximización de la correlación entre las estructuras teórica y real. Sin embargo, dicho procedimiento no resulta, en general, correcto dado que la existencia de una alta correlación entre las estructuras, no supone, necesariamente, que ambas sean idénticas, sino sólo que se comportan igual. El coeficiente de correlación mide la fuerza de la relación entre dos variables, y el sentido de la misma, pero no la coincidencia o concordancia del valor de sus observaciones. De hecho, un cambio en las unidades de medida de una de las magnitudes no afectaría a la correlación, aunque si indudablemente a la concordancia.

En otro orden de cosas, tampoco el método empleado en la resolución del problema de maximización resulta idóneo. El algoritmo iterativo aplicado- propuesto por Fletcher y Powell-, introduce variabilidad en sus resultados según cuál sea el punto de partida considerado, lo cual cuestiona la unicidad y representatividad de la solución ofrecida al existir la posibilidad de presentar como solución un máximo que sólo sea local, y no el global de la función (Everitt, 1987).

La conjunción de una medida de partida potencialmente confusa junto con las limitaciones del procedimiento de resolución aplicado, plantean la necesidad de

afrontar una reforma de la metodología en una doble vertiente, estimación de coeficientes y análisis de bondad, cuestiones abordadas en los siguientes epígrafes.

3. Proceso de estimación

Con la finalidad de intentar resolver el primer problema señalado, proponemos emplear medidas derivadas de la teoría de la información con el objetivo de estimar la estructura teórica que presente mínima divergencia con la estructura real objeto de estudio.

Una de las medidas pioneras de la teoría de la información es la denominada entropía de Shannon, la cual junto con otras medidas similares, han sido empleadas como indicadores de la diversidad existente dentro de una población[3].

La entropía de una distribución puede ser entendida como el desorden existente en la misma, es decir, la incertidumbre asociada a un determinado fenómeno. Dicha incertidumbre puede ser cuantificada a partir de un sistema de probabilidad. Así, si un fenómeno lleva asociado una distribución de probabilidad uniforme, presentará una mayor carga de incertidumbre que cualquier otro.

A partir de la medida de entropía propuesta por Shannon, Kullback y Leibler (1951) introducen una medida de divergencia entre dos distribuciones. Consideremos dos variables aleatorias discretas X e Y cuyas probabilidades asociadas son $\mathbf{P} = \{p_1,...,p_n\}$ y $\mathbf{Q} = \{q_1,...,q_n\}$, respectivamente. Se define la distancia de Kullback-Leibler como la expresión que sigue:

$$D(P,Q) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right)$$

Siguiendo la línea desarrollada por Golan, Judge y Robinson (1994) se plantea la siguiente expresión a partir de la divergencia anterior de Kullback y Leibler, consideradas dos distribuciones de probabilidad conjuntas recogidas en $\mathbf{P} = \{p_{ij}\}$ y $\mathbf{Q} = \{q_{ij}\} \forall i, j = 1,...,n$:

[3] Ver Rao (1982).

$$\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Dicha medida permite detectar la divergencia entre dos distribuciones bidimensionales discretas. Además, ofrece una alternativa a la estimación de la matriz teórica (δ_{ij}) que suponga la mínima divergencia con la estructura real objeto de estudio (x_{ij}).

La especificación del modelo centro *versus* periferia, a partir de la delimitación de un conglomerado de nodos que constituya un núcleo generador de actividad, se inspira en el principio de mínima entropía sujeto a una serie de restricciones:

$$\text{Min} \sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

$$\text{s.a.} \sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij} = 1$$

$$0 \le p_{ij} \le 1$$

$$\sum_{j=1}^{n} p_{ij}\delta_{.j} = \delta_{i.}$$

Donde p_{ij} y q_{ij} corresponden respectivamente a las proporciones de conexiones teóricas (δ_{ij}) y observadas (x_{ij}) en las estructuras consiguientes recogidas en la Tabla Nº 1:

Tabla Nº 1. Organigrama de estructuras

RED TEÓRICA			
	1	...	n
1	δ_{11}	...	δ_{1n}
...	...	...	...
n	δ_{n1}	...	δ_{nn}

RED REAL			
	1	...	n
1	x_{11}	...	x_{1n}
...	...	...	...
n	x_{n1}	...	x_{nn}

de tal forma que:

$$p_{ij} = \frac{\delta_{ij}}{\sum_{i=1}^{n}\delta_{ij}} = \frac{\delta_{ij}}{\delta_{.j}} \qquad q_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}} = \frac{x_{ij}}{x_{.j}}$$

La solución al proceso de minimización planteado se obtiene utilizando el método de los multiplicadores de Lagrange[4], cuya expresión es (Golan, et. al. 1994):

$$p_{ij} = \frac{q_{ij}}{\Omega_j(\lambda_i)} \exp\left[-\lambda_i \delta_{.j}\right]$$

Donde

$$\Omega_j(\lambda_i) = \sum_{i=1}^{n} q_{ij} \exp\left[-\lambda_i \delta_{.j}\right]$$

y λ_i representa los multiplicadores de Lagrange asociados a las restricciones.

Es interesante destacar algunas de las propiedades del proceso de minimización generado a partir de la entropía cruzada[5], tales como la unicidad de la solución, que muestran que este procedimiento es potencialmente más adecuado que el algoritmo empleado por Borgatti y Everett el cual no permite garantizar la representatividad de la solución[6]. El método planteado permite además ofrecer estructuras similares en las cuales se conserven tanto los ceros como la intensidad de las variables. Los coeficientes q_{ij} importantes o relativamente elevados son representados por p_{ij} también relativamente considerables en la matriz teórica y, los coeficientes que son nulos o positivos en la matriz observada, seguirán el mismo patrón en la estructura imagen[7]. La posible inclusión además de ecuaciones no lineales, junto con desigualdades supone una ventaja operacional de la que otros métodos menos rápidos y eficientes carecen. Este rasgo permite la inclusión de información adicional derivada de otras fuentes o estudios, en la medida en la que sea posible su definición a partir de apropiadas ecuaciones.

Dado que nuestro interés es la determinación de los índices de centralidad (c_i), para diferenciar los actores centrales de los periféricos, se debe proceder a su estimación a partir de los resultados obtenidos anteriormente (p_{ij}).

---

[4] Dadas las características del problema, éste ha de se resuelto numéricamente.
[5] Se pueden consultar dichas propiedades en Blien (2001).
[6] Ver Everitt (1987), Cover y Thomas (1991) para una revisión de ambos métodos respectivamente.
[7] Ver Blien y Graef (1997).

Puesto que por construcción $p_{ij}=\frac{\delta_{ij}}{\sum_{i=1}^{n}\delta_{ij}}=\frac{\delta_{ij}}{\delta_{.j}}$, las conexiones teóricas entre un par de actores i-*ésimo* y j-*ésimo* se pueden determinar un sistema de ecuaciones como el que sigue:

$$\delta_{ij} = p_{ij}\sigma_{.j}$$

$$\delta_{.j} = \sum_{i=1}^{n}\delta_{ij}$$

A partir del cual se establecen los índices de centralidad como[8]:

$$\delta_{ij} = c_i c_j \quad \forall i, j = 1,...,n$$

$$c_i \geq 0;\ c_j \geq 0$$

4. Análisis de bondad

La adaptación de la matriz teórica a la matriz observada, constituye un punto de análisis interesante y posterior a la determinación de los datos necesarios en nuestro modelo. Se presentan, para ello, dos posibles líneas de trabajo con el objetivo de estudiar la fiabilidad del modelo planteado.

La primera mantiene el hilo conductor derivado de la teoría de la información. La distancia de Kullback y Leibler sobre las variables bidimensional supone una medida del grado de exactitud entre las estructuras consideradas. La segunda vía de estudio ofrece un análisis gráfico del nivel de concordancia existente entre las redes planteadas a partir de la definición de límites de concordancia a un nivel de confianza $(1-\alpha)$.

4.1. Divergencia distributiva

Una medida de divergencia directa extensamente aplicada por sus adecuadas propiedades es la distancia de Kullback y Leibler (1951), empleada ya en el proceso de estimación.

[8] Correspondería a un sistema de ecuaciones sobreidentificado.

Dadas dos distribuciones de probabilidad $\mathbf{p} = [p_1 ... p_n]$ y $\mathbf{q} = [q_1 ... q_n]$, un indicador de la divergencia entre dos distribuciones de probabilidad puede ser entendido como una medida del error cometido al considerar que la distribución correcta es **q**. Su expresión es la que sigue:

$$D(\mathbf{p} \mid \mid \mathbf{q}) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}$$

Muestra, por tanto, la ineficiencia o el error cometido al asumir que la distribución correcta es **q** cuando en realidad lo es la distribución desconocida **p.**

Medida acotada inferiormente por cero $(D(\mathbf{p} \mid \mid \mathbf{q}) \geq 0)$, alcanzará dicho valor, si y sólo si, **p**=**q**. Y a medida que la divergencia entre las distribuciones sea mayor, dicho indicador crecerá[9]. Además debe considerarse que $D(\mathbf{p} \mid \mid \mathbf{q}) \neq D(\mathbf{q} \mid \mid \mathbf{p})$, lo cual resulta un rasgo relevante en la decisión de la matriz sobre la cual se realiza la comparación.

Considérese ahora dos distribuciones de probabilidad bidimensionales $\mathbf{P} = \{p_{ij}\}$ y $\mathbf{Q} = \{q_{ij}\}$ $\forall i = 1,...,n; \forall j = 1,...,n$, cuyos elementos p_{ij} y q_{ij}, definidos anteriormente, corresponden respectivamente a las proporciones de relaciones teóricas (δ_{ij}) y observadas (x_{ij}).

La distancia de Kullback y Leibler (1951) asociada a estas variables bidimensionales, con comportamiento análogo al anteriormente expresado para una variable unidimensional, ofrece un indicador de la concordancia existente entre la estructura teórica y real:

$$D(\mathbf{p}_{xy} \mid \mid \mathbf{q}_{xy}) = \sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

[9] Alcanzará el valor log(n) en el caso en que una de las distribuciones sea degenerada y la otra presente un reparto uniforme.

4.2. El método de altam-bland

El método desarrollado por Altam-Bland (1986) permite visualizar gráficamente el nivel de concordancia existente entre las dos estructuras consideradas.

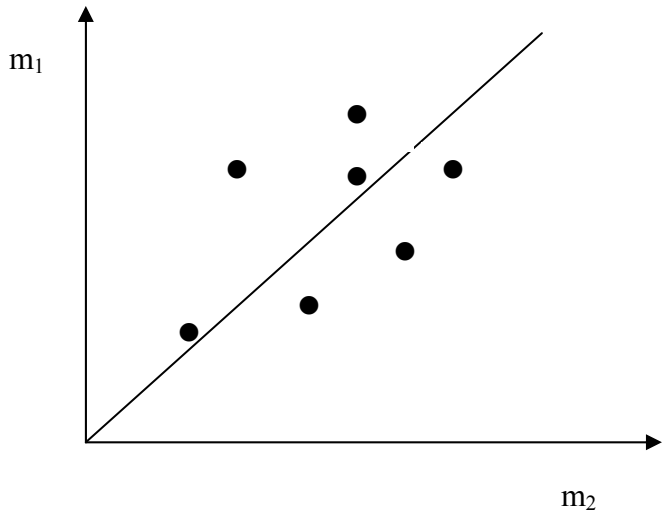
Sea m_{ij}, $(\forall i = 1,...,n;\ j = 1,2)$ la relación i-*ésima*[10] especificada en la matriz j-*ésima*, la información disponible podrá ser resumida en una tabla de doble entrada como la expuesta a continuación, donde el tamaño de muestra (n) para cada una de las dos situaciones experimentales- tablas real y teórica- coincide.

Tabla Nº 2

	Red	
Observaciones	**Observada**	**Teórica**
1	m_{11}	m_{12}
...	...	...
i	m_{i1}	m_{i2}
...	...	...
n	m_{n1}	m_{n2}

La representación de la nube de puntos de las observaciones consideradas junto a la bisectriz del primer cuadrante, supone una aproximación inicial e intuitiva al grado de concordancia existente. De forma que, cuanto mayor sea la proporción de datos sobre la línea mencionada, menor será la discrepancia existente entre los resultados estimados y observados.

Gráfico Nº 1



[10] Sea considerada cada celda de la tabla analizada como una observación i-*ésima*, orden que será mantenido para el estudio de las restantes matrices.

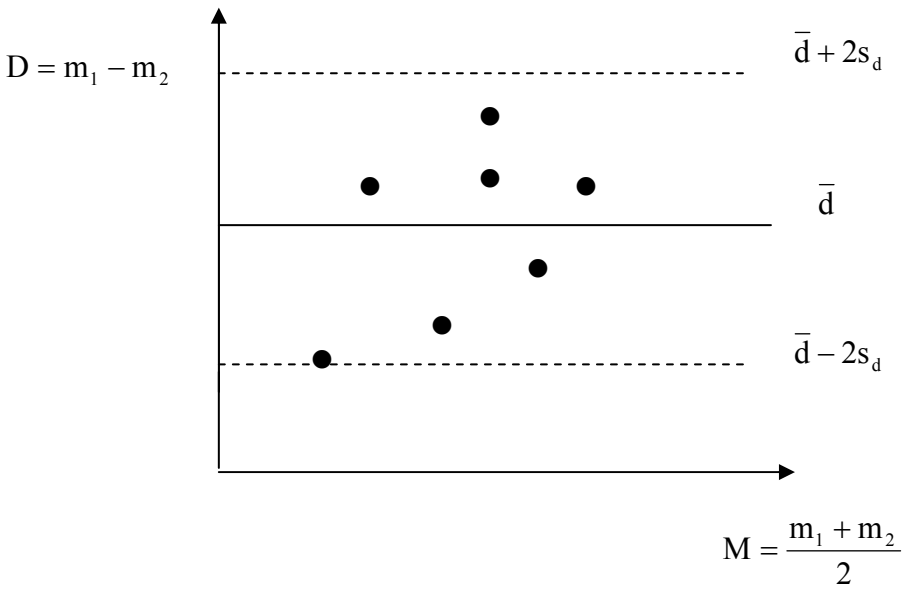
Donde m_1 y m_2 representan los datos relacionales de la red observada y teórica respectivamente.

Dado que, previsiblemente, ambas estructuras divergirán en el grado de sus interrelaciones, resulta interesante medir el grado de concordancia alcanzado.

El coeficiente de correlación usualmente empleado con este fin no resulta, como ya se ha comentado, adecuado por lo que la siguiente representación supone una alternativa sencilla que proporciona dos límites aleatorios entre los que con cierta probabilidad o confianza $(1-\alpha)$, se halle comprendido el nivel de divergencia existente entre las estructuras analizadas.

Sea D la variable diferencia resultante de la comparación de la estructuras de las dos redes, y M la media de las observaciones respectivas de las mismas. La gráfica de la nube de puntos que representa a las variables media y diferencia se representa en el gráfico Nº 2.

Gráfico Nº 2



La delimitación de una banda en la cual se sitúen discrepancias pequeñas entre ambas redes se puede generar como $(\bar{d} \pm 2s_d)$. De hecho, si la variable diferencia

se distribuyese normalmente, aproximadamente el 95% de las mismas se situarían dentro de los límites establecidos como $\left(\bar{d} \pm 1{,}96 s_d\right)$.

Si las divergencias observadas entre las dos estructuras se sitúan dentro de los extremos establecidos, denominados límites de concordancia, no resultarán significativas y se puede considerar las redes como potencialmente intercambiables dada su similitud. En caso de que no haya un error sistemático los puntos se distribuirán de forma aleatoria a uno y otro lado de la recta que represente la diferencia nula[11]. Dado que estos límites son estimaciones puntuales aplicadas al conjunto de la población, deben ser complementadas con sus intervalos de confianza respectivos[12], las bandas aleatorias a las cuales pertenecerán los parámetros desconocidos con cierto nivel de confianza $\left(1-\alpha\right)$ bajo el supuesto de normalidad[13]:

$$IC_{\text{lim inferior concordancia}}\left[\bar{d} - 2s_d \pm k\sqrt{\frac{3s_d}{n}}\right] \qquad IC_{\text{lim superior concordancia}}\left[\bar{d} + 2s_d \pm k\sqrt{\frac{3s_d}{n}}\right]$$

[11] Las relaciones entre ambas variables puede mostrar la existencia de cierto sesgo, en cuyo caso, los datos deben ser sometidos a ciertos tratamientos estadísticos. La aplicación de logaritmos, puede solventar el problema en aquellos casos más sencillos. En caso contrario se requiere la aplicación de una combinación de métodos de regresión que establezcan el intervalo de confianza establecido. Ver Bland y Altam (1999).

[12] La varianza del intervalo de confianza inicial, bajo el supuesto de independencia entre $\bar{d}$ y s_d, responde a:

$$\text{Var}\left(\bar{d} \pm 1{,}96 s_d\right) = \text{Var}\left(\bar{d}\right) + 1{,}96^2\,\text{Var}\left(s_d\right) = \frac{s_d^2}{n} + 1{,}96^2\,\frac{s_d^2}{2(n-1)} = \left(\frac{1}{n} + \frac{1{,}96^2}{2(n-1)}\right)s_d^2$$

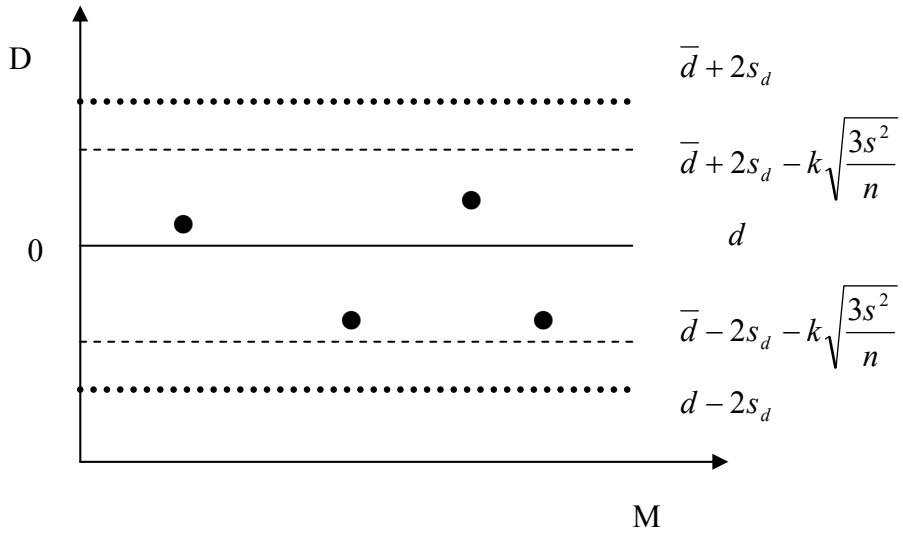
Expresión que puede ser aproximada para tamaños de muestra elevados por:

$$\text{Var}\left(\bar{d} \pm 1{,}96 s_d\right) = (2{,}92)\frac{s_d^2}{n} \cong 3\frac{s_d^2}{n}$$

[13] Si las variables no cumplen el supuesto de normalidad, y existen fuertes discrepancias entre las redes consideradas, la estimación de los intervalos se debe plantear a partir de métodos no paramétricos. Dada la proporción de diferencias mayores a cierto valor de referencia, o un determinado centil para, por ejemplo el 10%, la banda de confianza se construye aplicando el modelo binomial para la proporción o el error estándar del centil. Aún así estos métodos resultan menos exactos o confiables que aquellos que emplean el supuesto de normalidad (Bland y Altman, 1999).

La representación gráfica donde se recogen estos límites aparece en el gráfico siguiente, tal que la línea pespunteada gruesa representa la banda de concordancia y sus límites inferiores de confianza son delimitados por pespuntes finos.

Gráfico Nº 3



5. Conclusiones

Los modelos centro-periferia desarrollados por Borgatti y Everett (1999), constituyen una herramienta de análisis de valiosa utilidad, que son capaces de modelizar la idea clásica de la existencia de una estructura constituida por un núcleo activo, formado por un entramado denso y compacto de actores, frente a un conglomerado disperso en sus relaciones y poco conectado.

El proceso de estimación aplicado por estos autores presenta sin embargo, una serie de restricciones. La conjunción de una medida de partida potencialmente confusa junto con las limitaciones del procedimiento de resolución, plantean la necesidad de afrontar una reforma de la metodología en una doble vertiente, estimación de coeficientes y análisis de bondad, cuestiones abordadas en el presenta trabajo.

6. Bibliografía

Bland, J. Martin, Altam, Douglas G. (1986). "Statistical methods for assessing agreement between two methods of clinical measurement". *Lancet*, i, pp.307-310.

Bland, J. Martin, Altam, Douglas G. (1999). "Measuring agreement in method comparison studies". *Statistical Methods in Medical Research*, Nº 8, pp. 135-160.

Blien, Uwe, Graef, Friedrich (1997). "Entropy optimising methods for the estimation of tables". *Proceedings of the 21st Annual Conference of the Gesellschaft für Klassifikation e.V.*, University of Potsdam, March 12-14.

Blien, Uwe, Tassinopoulos, Alexandros (2001). "Forecasting regional employment with the ENTROP method". *Regional Studies*, Vol. 35, Nº 2, pp.113-124.

Borgatti, Stephen P., Everett, Martin G. (1999). "Models of Core/Periphery Structures". *Social Networks*, Nº 21, pp. 375-395.

Cover, Thomas M., Thomas, Joy A. (1991). *Elements of Information Theory*. United States of America: Wiley series in telecommunications.

Everitt, Brian (1987). *Introduction to optimisation methods and their application in statistics*. New York: Chapman and Hall Ltd,.

Golan Amos, et. al. (1994). "Recovering information from incomplete or partial multisectoral economic data". *The Review of Economics and Statistics,* Nº 76, pp. 541-549.

Kullback, Solomon, Leibler, Richard A. (1951). "On information and sufficiency", *Annals of Mathematical Statistics*, Nº 22, pp. 79-86.

Lin, Lawrence, Hedayat, A.S., Sinha, Bikas, Yang, Min (2002). "Statistical Methods in Assessing Agreement: Models, Issues, and Tools". *Journal of the American Statistical Association*, Vol. 97, Nº 457, pp. 257-270.

Lin, Lawrence (1989). "A concordance correlation coefficient to evaluate reproducibility". *Biometrics*, Nº 45, pp.255-268.

Rao, C. Radhakrishna (1982). "Diversity and dissimilarity coefficients: A unified approach". *Theoretical Population Biology*, Vol. 25, pp. 24-43.

Shannon, Claude E. (1948) "A mathematical theory of communication". *Bell Systems Tech Journal*, Vol. 27, pp. 379-423, 623-659.