

USO DEL ALGORITMO DE KOHONEN, APLICADO AL ESTUDIO DE LA LOCALIZACIÓN Y ACCESIBILIDAD DE REVISTAS CIENTÍFICAS EN BIBLIOTECAS UNIVERSITARIAS.

- Autores:** María J. Reyes Barragán
mjreyes@alcazaba.unex.es
Vicente Guerrero Bote
vicente@alcazaba.unex.es
Felipe Zapico Alonso
fzapalo@alcazaba.unex.es
Facultad de Biblioteconomía y Documentación. Universidad de Extremadura
- Resumen:** Tradicionalmente las revistas científicas utilizadas por los investigadores universitarios se han encontrado ubicadas físicamente en la sede de los distintos Departamentos, lo que ha producido la dispersión de títulos y la variación de condiciones en cuanto a su accesibilidad. El objeto del presente trabajo es racionalizar la accesibilidad a la información disponible. Con este fin hemos analizado los usos de revistas por los distintos departamentos basándonos en las citas recibidas en los trabajos realizados (ponderadas por el número de autores de cada trabajo). Para este análisis hemos utilizado el algoritmo de Kohonen, un modelo de red neuronal artificial, capaz de clasificar las entradas (las revistas). La ventaja que tiene este método sobre otros de reducción de la dimensión radica en la posibilidad de clasificar y representar un gran número de revistas junto a los departamentos en un espacio bidimensional.
- Palabras clave:** Algoritmo de Kohonen ; Bibliotecas universitarias; revistas científicas; análisis de citas; Clusters; Bibliometría; Producción científica; Redes neuronales
- Abstract:** The scientific journals that university researchers use have traditionally been located physically inside the various Departments. This has led to the dispersion of titles and the variability of conditions of accessibility. The goal of the present work is to rationalize accessibility to the available information. To this end, we analysed the uses of journals by the different departments on the basis of the citations received in their publications (weighted by the number of authors in each work). For this analysis we used Kohonen's algorithm, which is an artificial neural network capable of classifying inputs (in this case the journals). The advantage of this dimension-reducing method over others lies in the possibility of classifying and representing a great number of journals together with the departments in a two-dimensional space.
- Keywords:** Kohonen's algorithm; university libraries; scientific journals; citation analysis; clusters; bibliometrics; scientific output; neural networks

1. Introducción

Las revistas científicas, hoy por hoy, son las herramientas básicas con que cuentan los científicos como medio inmediato de transmisión de

conocimiento y generador de otros nuevos que a su vez volverán a ser transmitidos.

Tradicionalmente las revistas científicas utilizadas por los investigadores universitarios se han encontrado ubicadas físicamente en los distintos departamentos, lo que ha producido la dispersión de títulos y distintas condiciones en cuanto a su accesibilidad. El objeto del presente trabajo es racionalizar la accesibilidad a la información disponible, para poder establecer los mecanismos y filtros necesarios para que los límites entre lo disponible y lo idóneo tiendan a unificarse.

Han sido muchos los criterios utilizados para evaluar las revistas científicas, su concepción y ejecución, aunque tendentes a un mismo fin ofrecen ciertas diferencias basadas predominantemente en la incidencia relativa, mayor o menor, de factores subjetivos, en relación con los criterios objetivos. Nuestro análisis ha partido principalmente de criterios objetivos, basándonos en el método objetivo del análisis de citas como propone E. Jiménez Contreras (1994), que las considera como una expresión del uso de los fondos. En este sentido su opción concuerda con la mantenida por (Altuna y Lancaster, 1992, Swuigger y Wilkes 1991, Bookstein 1988, Gardfield,1972), si bien tenemos que hacer la salvedad del sesgo previo que realizan las bases de datos, pero que son, sin embargo, las únicas herramientas disponibles para cuantificar.

Como base para poder realizar un mapa topológico de las revistas científicas más usadas se ha utilizado el algoritmo de kohonen, un modelo de red neuronal artificial capaz de clasificar las revistas (entradas) junto a los departamentos en un espacio bidimensional.

2. Metodología

En base a nuestros objetivos de racionalizar el acceso a la información y detectar que revistas son más utilizadas y quienes las utilizan fue necesario la aplicación práctica de principios bibliométricos. El estudio se realizó en una institución universitaria (Universidad de Extremadura), en las áreas de ciencia y tecnología, en un período de tiempo de dos años (1996-1997).

Podemos decir que nuestro análisis se sustenta en dos pilares: el análisis de citas y los mapas auto-organizativos de Kohonen. Sobre el primero tendríamos que decir que el análisis de citas, en la actualidad, parece ser el método más pragmático y fiable para el desarrollo de colecciones de revistas científicas, aunque ningún método "per se" es totalmente válido. Esto hizo que se incorporaran otros factores que podrían proporcionar un valor añadido al estudio. Expondremos brevemente los pasos que tuvieron que darse:

- 1.- Obtener la producción científica de la Universidad en las áreas de ciencia y tecnología de la base de datos del SCI (Science Citation Index)
- 2.- Identificar los autores de los trabajos con las distintas unidades departamentales
- 3.- Realizar el análisis de las revistas más citadas a partir de las referencias bibliográficas contenidas en los trabajos y poder determinar el corpus de revistas más citadas
- 4.- Aplicar la ley de Bradford para delimitar el núcleo y la primera zona de las revistas más citadas

5.- Determinar el uso de revistas por departamentos. Para ello las revistas que aparecían citadas en un trabajo se ponderaron por el número de firmantes del trabajo y los autores habían sido identificados previamente con una unidad departamental.

6.- Asignar el factor de Impacto a todas las revistas citadas con una frecuencia superior a 10 usos (5 citas).

7.- Fijar la disponibilidad en bibliotecas para determinar hasta que punto la dispersión de las revistas científicas más usadas guardan relación con la disponibilidad que tienen los usuarios a estas.

En el segundo pilar se utilizó el algoritmo de Kohonen, que como exponemos a continuación es un modelo de red neuronal artificial, capaz de clasificar las entradas (revistas) sobre una rejilla, de modo que en cada nodo de la rejilla se forma un cluster con revistas que tienen un perfil de uso similar, y determinar el comportamiento de las unidades departamentales en función de los usos a revistas. Esto nos permite tener una visión topológica de las revistas clasificadas en función de los usos realizados por los departamentos. También nos va a permitir representar topológicamente las revistas en función de los parámetros analizados anteriormente (revistas más usadas, factor de impacto, disponibilidad en biblioteca, etc.)

Para utilizar este modelo, al igual que otros, se requiere representar las revistas (entradas) vectorialmente. Se utilizan vectores de veintiuna componentes que ponderan los usos de las revistas de los veintidós departamentos.

La ventaja que tiene este método sobre otros de reducción de la dimensión radica en la posibilidad de clasificar un gran número de revistas junto a los departamentos en una rejilla bidimensional, ofreciendo así una organización topológica que aportara mucha información sobre los usos de cada una de las revistas, incluso para determinar la localización de cada una de éstas (teniendo en cuenta que lo que más se usa debe estar lo más cerca posible).

3. Algoritmo de Kohonen

A pesar de la enorme complejidad de la corteza cerebral desde el punto de vista microscópico, a escala macroscópica tiene una estructura uniforme, incluso al pasar de un cerebro a otro. Los centros correspondientes a actividades concretas como el pensamiento, visión, oído, funciones motoras, etc., yacen en zonas concretas de la corteza y éstas se ubican de una cierta manera con respecto a las demás. Un ejemplo es el denominado mapa tonotópico de las regiones auditivas, en el cual las neuronas próximas entre sí responden a frecuencias de sonido similares. Otro ejemplo es el mapa somatotópico que está representado artísticamente mediante el conocido homúnculo.

La corteza es una capa extensa (de aproximadamente 1 m^2) y fina (entre 2 y 4 mm de grosor) que consta de seis capas de neuronas de distintos tipos. Está plegada para maximizar la superficie que cabe en el interior del cráneo, no obstante, para nuestro interés es como si fuera una superficie.

Es posible que en gran medida este mapa esté predestinado por la constitución genética. No obstante, el interés por descubrir como podía llegarse a una organización de este tipo, fue lo que condujo a Kohonen a investigar sobre el tema (Kohonen, T 1982,1989, 1995). El producto de estas investigaciones ha sido el modelo de redes neuronales artificiales que lleva su nombre, las cuales son capaces de organizar topológicamente las entradas.

En realidad, es un modelo competitivo muy similar al de contrapropagación con un tipo de capas competitivas ligeramente diferente a las de *centro activo/periferia inactiva*. De esta forma, se puede decir que las redes de Kohonen no son más que una capa competitiva que puede utilizarse formando parte de redes como las anteriores o independientemente.

La principal diferencia viene en la influencia que una neurona tiene sobre sus vecinas. En la contrapropagación, cada neurona tenía una realimentación positiva, mientras que influía negativamente en el resto de las neuronas de la misma capa. En este caso la influencia que cada neurona ejerce sobre el resto de las neuronas de su capa va a ser función de la distancia entre las mismas. La función más utilizada para esto es la conocida función *sombrero mejicano* que podemos ver en la figura 1(a) (que normalmente se aproxima mediante la función de la figura 1(b)). Es decir, cada neurona ejerce una influencia positiva sobre sí misma y sobre las neuronas topológicamente cercanas. Esta influencia va decreciendo a medida que aumenta la distancia entre las neuronas, hasta hacerse negativa, para tener finalmente una influencia positiva sobre las más alejadas.

Esto tiene una base biológica, ya que se ha comprobado que en determinados primates se producen interacciones laterales de tipo excitatorio entre neuronas próximas en un radio de 50 a 100 micras, de tipo inhibitorio en una corona circular de 150 a 400 micras de anchura alrededor del círculo anterior, y de tipo excitatorio muy débil, prácticamente nulo, desde ese punto hasta una distancia de varios centímetros (Hilera y Martínez, 1995).

Como consecuencia de esto en la capa se da una burbuja de actividad, formada por todas aquellas unidades que están cercanas a la ganadora, las cuales participan del refuerzo correspondiente al aprendizaje. Fruto de ello los *mapas autoorganizativos de Kohonen* (a diferencia de otras redes) dan lugar a una *correspondencia que respeta la topología* entre los datos de entrada y las unidades competitivas. En estos mapas, unidades de la capa oculta cercanas físicamente responden a vectores de entrada que se encuentran igualmente próximos. Estas neuronas se suelen organizar en una rejilla bidimensional.

La simulación hardware conlleva la creación de una capa competitiva de cierta complejidad, sin embargo, el proceso que lleva a cabo durante el aprendizaje cada vez que se le presenta un vector, lo podemos resumir en los siguientes pasos:

- *Seleccionar como nodo ganador (representado por un vector de pesos de la misma dimensión de entrada) al más cercano al vector presentado.*
- *Ajustar los vectores de pesos del nodo ganador y de los correspondientes a su vecindad acercándolos al de entrada (en algunos casos el refuerzo es igual para toda la vecindad y en otros decrece a medida que aumenta la distancia al ganador).*

En el entrenamiento se le presentan repetidamente los vectores seleccionados para tal fin, en un orden aleatorio, a la vez que se reduce progresivamente la vecindad y el parámetro de aprendizaje (que marca la cantidad que se desplazan los vectores de pesos para ajustarse a las entradas) para forzar la estabilidad de la red. Tras esta fase se llega a una configuración en la que las neuronas topológicamente cercanas en la red resultan ganadoras ante cúmulos de vectores cercanos en el espacio de entrada. Esto es debido a que han participado juntas de muchos refuerzos. En algunas ocasiones a este algoritmo se le añade una conciencia que favorece que resulten vencedoras aquellas neuronas que han ganado menos competiciones, con el fin de que las victorias se repartan por toda la red.

Debido a esta organización topológica a veces lo único que interesa es el clustering llevado a cabo por la capa oculta, y se selecciona todo el conjunto de vectores para el entrenamiento con el único motivo de ver la organización topológica resultante. Dentro de esta última aplicación existen dos posibilidades, por un lado si se tienen más unidades ocultas que vectores de entrenamiento lo que se consigue es una proyección óptima sobre la topología que se elija. Si el número de unidades es menor al número de vectores lo que se consigue es una capa que hace clustering y ordena cada cluster topológicamente. El número de clusters resultantes será igual al de neuronas que formen la capa oculta.

Así podemos decir que de una forma iterativa, pero, sencilla consigue no sólo un buen análisis de cluster sino una buena organización topológica. Sin embargo, al igual que otros algoritmos, tiene algunas características que no satisfacen mucho desde el punto de vista matemático como son: la terminación forzada por el número de iteraciones, la convergencia no garantizada, la dependencia del orden de entrada de los datos, la obtención de la estabilidad con la reducción de la velocidad de aprendizaje, la generación de una partición clásica en lugar de difusa, etc. Ha habido intentos de mejora y de fuzzificación, como el *Fuzzy Kohonen Clustering Networks* (Bezdek 1992), donde se intenta incorporar algunas de las características del método de las *c-medias*, que proporciona una salida difusa. En todos estos intentos se producen ciertas mejoras, sin embargo, todos suelen tener el denominador común de la pérdida de la organización topológica que caracteriza a este algoritmo.

Recientemente se han utilizado redes de Kohonen para la generación de *mapas topológicos* de un conjunto de documentos, etiquetando incluso las zonas de influencia de cada palabra o término (Honkela et al., 1995; Lin 1997; Kohonen et al., 1999a, 1999b, Guerrero et al., 2001).

4. Resultados

Se obtuvieron del SCI 375 trabajos (88% artículos) que suponen un 47 % de los trabajos publicados en esos años (Fuente: Memorias de investigación de la Universidad).

En la figura 2 aparecen representados las distintas unidades departamentales. Los puntos negros son neuronas o nodos de la rejilla, donde han sido clasificadas las revistas en función de los usos. Las gradaciones de color indican distancias entre las distintas neuronas, un indicativo de la proximidad temática de las revistas (clasificadas en las correspondientes neuronas). Una forma fácil de interpretarla es como mapas topográficos, donde los colores claros indican los valles, en los que las ciudades (neuronas) están mejor comunicadas y, por tanto, más relacionadas. Mientras que los colores oscuros trazan las montañas que son barreras de aislamiento. Una primera observación nos lleva a intuir que la mayor parte de los departamentos se ubican en valles donde se clasifican revistas que son de uso exclusivo de un departamento. Las zonas grises suelen corresponder con zonas de solapamiento entre departamentos, lo que se traduce en revistas compartidas por varios departamentos y por último las zonas oscuras denotan un aislamiento entre los clusters de departamentos.

A continuación (figura 3) se han representado las revistas más citadas en series de cinco, mediante una clave numérica asociada a títulos de revistas científicas y su disposición topológica. En una superposición de esta representación con la anterior observaríamos la disposición topológica de las revistas junto a las unidades departamentales, detectándose un solapamiento de áreas de conocimiento, traduciéndose en que son revistas usadas por distintos departamentos. Las revistas concentradas en los valles (zonas blancas) corresponden a revistas usadas específicamente por ese departamento, asociadas a áreas de conocimiento específicas. También aparecen revistas multidisciplinarias, pero, que son de mayor uso en ese clusters.

Las revistas más citadas determinadas por el núcleo y primera zona de la ley de Bradford se representan en la figura 4. Este lo conforman 181 revistas (7,34 % del total de revistas 2.465 títulos) que soportan el 54,46 % de las citas. Aparecen topológicamente las revistas junto a los departamentos en función de los usos, observándose que las revistas se concentran en los clusters (departamentos) más productivos y consolidados.

Seguidamente (figura 5) aparecen las revistas con mayor factor de impacto (6,5), situándose la mayoría de ellas en los clusters de las ciencias biomédicas, siendo en la mayoría de los casos revistas multidisciplinarias.

En cuanto a la disponibilidad en biblioteca se han representado en la figura 6 todas las revistas con una frecuencia de uso superior a 9. La disposición topológica advertida apunta hacia una desigual disponibilidad entre las distintas unidades departamentales, presentando mayor disponibilidad las ciencias químicas y las biomédicas.

La disponibilidad en núcleo y primera zona de la ley de Bradford (figura 7) se ha observado que prácticamente se concentra en dos departamentos (Ingenierías Químicas y Energéticas y Microbiología).

En general hallamos una desigual distribución de revistas disponibles en las distintas unidades departamentales y habría que aceptar como colección básica en las áreas de ciencia y tecnología, la determinada por el núcleo y primera zona de la ley de Bradford. Finalmente, teniendo en cuenta los resultados obtenidos, deberían tenerse en cuenta en la toma de decisiones para el desarrollo de la colección y plantearse la política de adquisiciones llevada a cabo en esta institución, dado el nivel de solapamiento entre áreas de conocimiento.

5. Conclusiones

El uso del algoritmo de Kohonen refleja fielmente los resultados obtenidos de otros métodos cuantitativos y cualitativos, adaptándose a la clasificación del conocimiento a través de revistas científicas.

La ventaja que proporciona este método sobre otros de reducción de la dimensión radica en la posibilidad de clasificar y representar gran número de revistas científicas, junto a las unidades departamentales en un espacio bidimensional. Ofrece así una organización topológica que aporta mucha más información sobre los usos de revistas científicas e incluso poder determinar la localización de cada una de estas. Refleja también los campos de conocimiento afines en función de los usos de revistas, traducándose en una similitud de necesidades de información.

Bibliografía:

- Altuna-Esteibar, B.; Lancaster, F. W. Ranking of journal in library and Information Science by research and teaching relatedness. *Serials Librarian*, 1992, n. 23, p. 1-10
- Bezdek, J. C. et al. Fuzzy kohonen clustering networks. *Proceedings first IEEE conference on fuzzy systems*. 1992, 1035-1043, San Diego.
- Bookstein, A. Applications of the bibliometrics distribution. En *Informetrics* 87/88. L. Egghe and R. Rousseau (eds.). Amsterdam: Elsevier, 1988.
- Garfield, Eugene. Citation analysis as a tool in journal evaluation. *Science*. 1972, vol. 178, n. 4060, p. 471-479.
- Guerrero Bote, V. P.; Moya Anegón, F. & Herrero Solana, V. (2001). Document Organization using Kohonen's Algorithm. *Information Processing & Management* (en prensa).
- Hilera, José R.; Martínez, Víctor J. *Redes neuronales artificiales, fundamentos, modelos y aplicaciones*. Madrid: Rama, 1995.
- Honkela, T; Pulkki, V; Kohonen, T. Contextual relations of words in grimm tales, analysed by self-organizing map. *Proceedings of international conference on artificial neural networks, icann-95*. 1995, 3-7, París.
- Jimenez-Contreras, E. et al. Determinación de las colecciones básicas de publicaciones periódicas en Hemerotecas científicas. En *actas de las IV Jornadas Españolas de Documentación Automatizada*, Gijón 1994. Oviedo: Universidad de Oviedo, 1994.

- Kohonen, T. Self-organized formation of topologically correct featuremaps. *Biological cybernetics* 1982, vol 43, 59-69. Reimpreso en el texto *neurocomputing* (J.Anderson y E.Rosenfeld eds.), M.i.t. press, 1988.
- Kohonen, T. *Self-organization and associative memory*. Berlin: Springer verlag, third edition, 1989.
- Kohonen, T. *Self-organization maps*. Berlin, Heidelberg: Springer verlag, 1995.
- Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., y Saarela, A. (1999a). Self organization of a massive text document collection. En Oja, E. and Kaski, S. (Eds.) *Kohonen Maps*, 171-182. Amsterdam: Elsevier.
- Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., y Saarela, A. (1999b). WEBSOM - A novel SOM-based approach to free-text mining. Neural Networks Research Centre (NNRC) at Helsinki University of Technology (HUT). <http://websom.hut.fi/websom/>.
- Lin, Xia. Maps displays for information retrieval. *Journal of the American Society for Information Science* 1997, vol 48, nº 1, p. 40-54.
- Swigger, Keith y Wilkes, Adeline. The use of citation data to evaluate serials subscriptions in academic library. *Serials review*. 1991, vol.17, n. 2, p. 41-52.

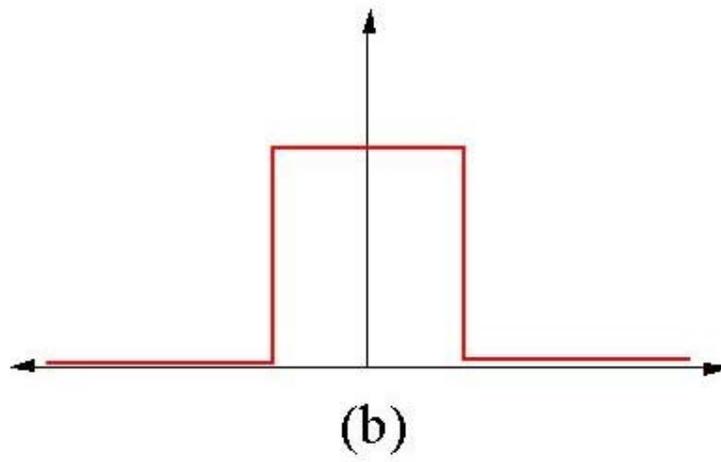
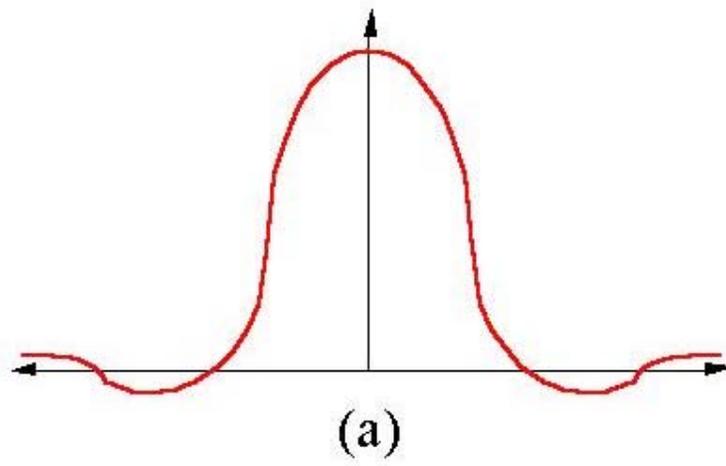


Figura 1

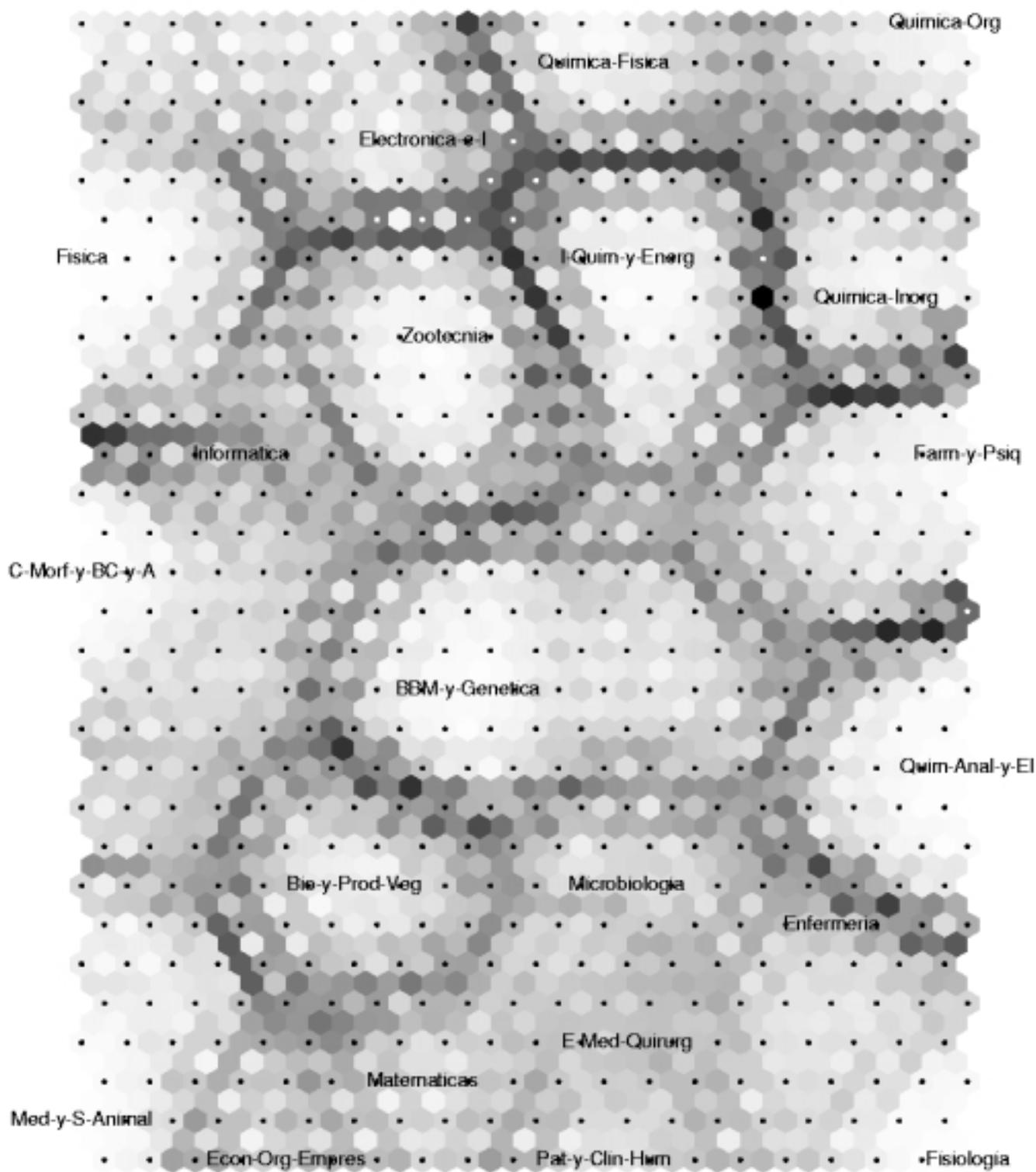


Figura 2

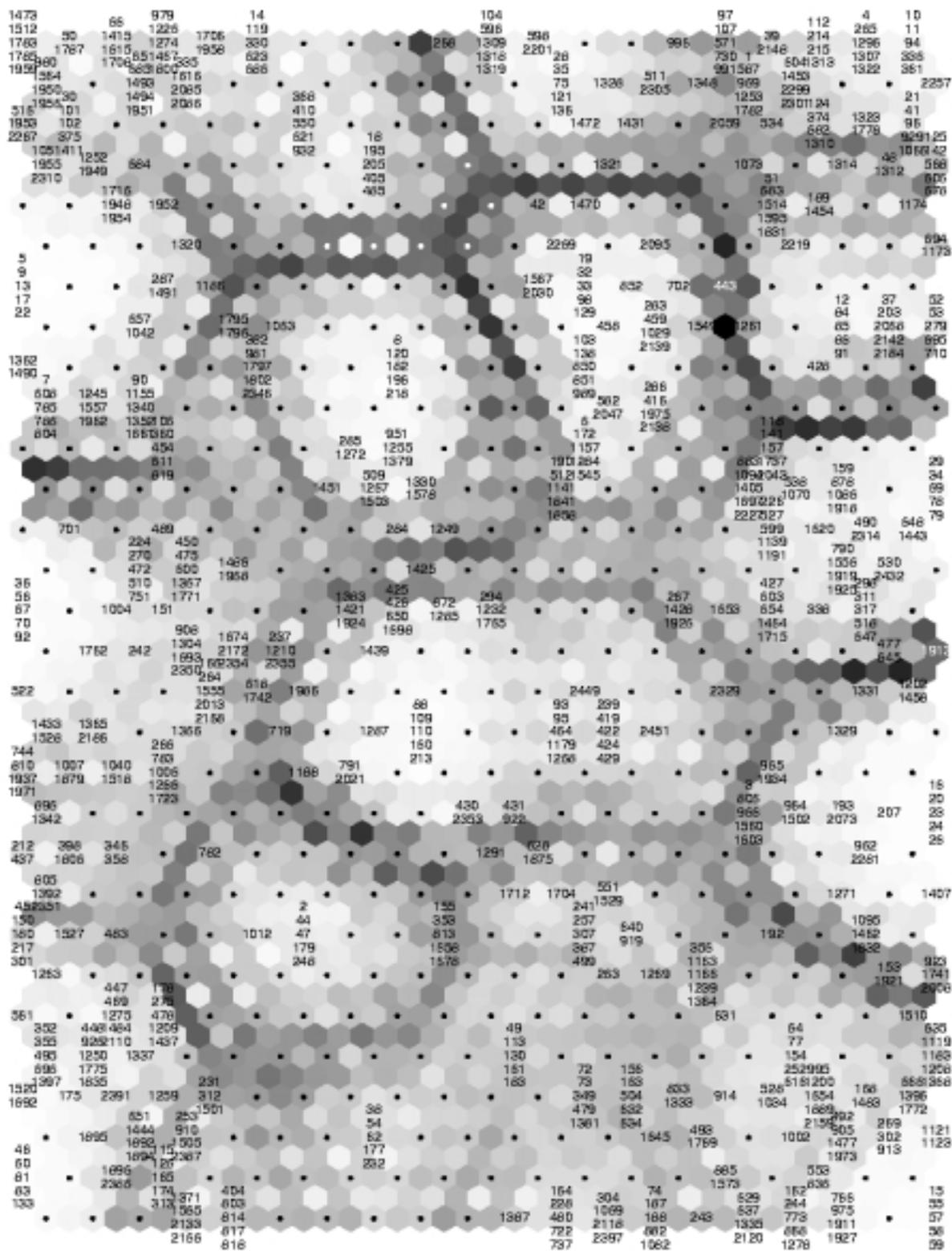


Figura 3

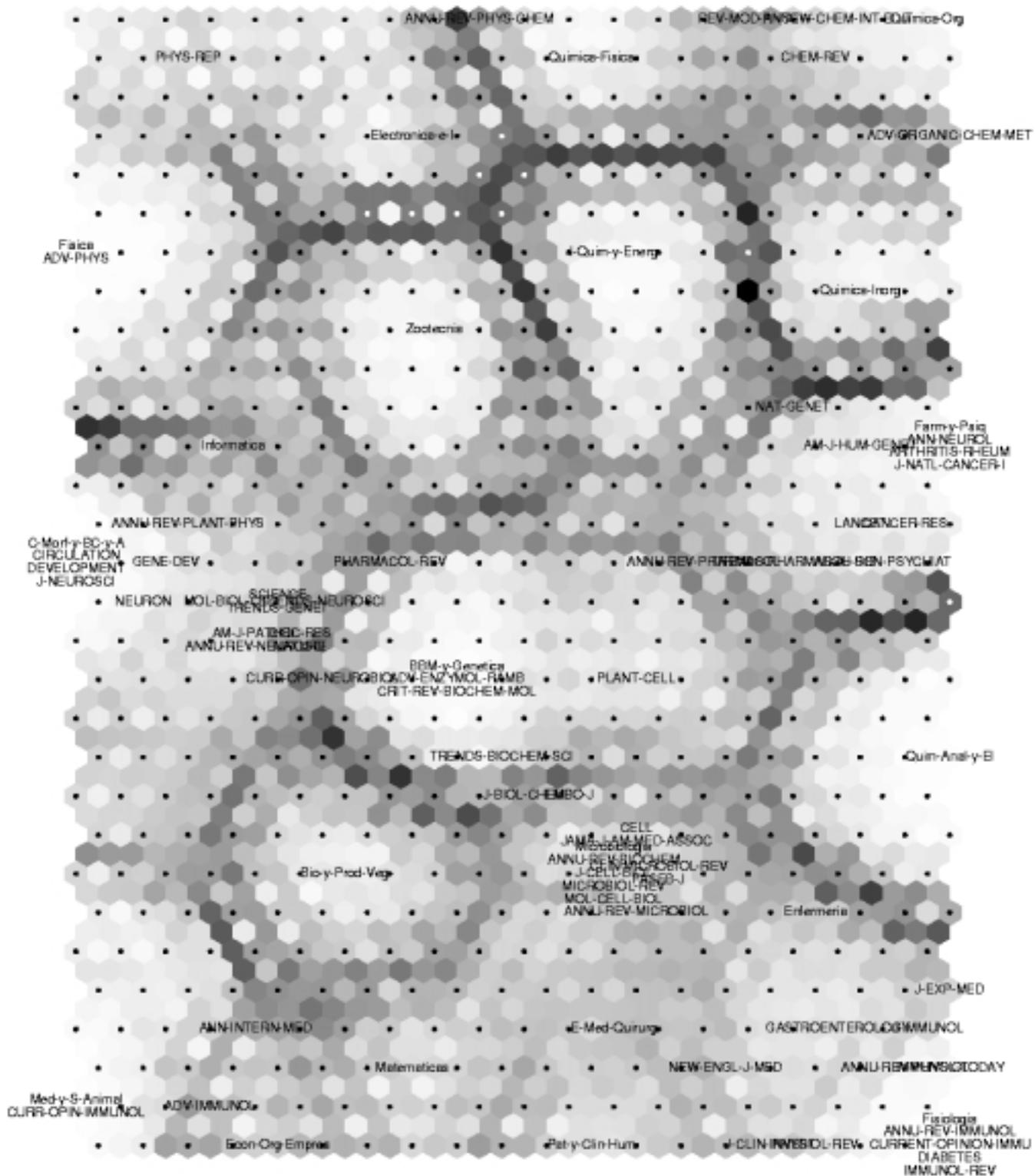


Figura 5

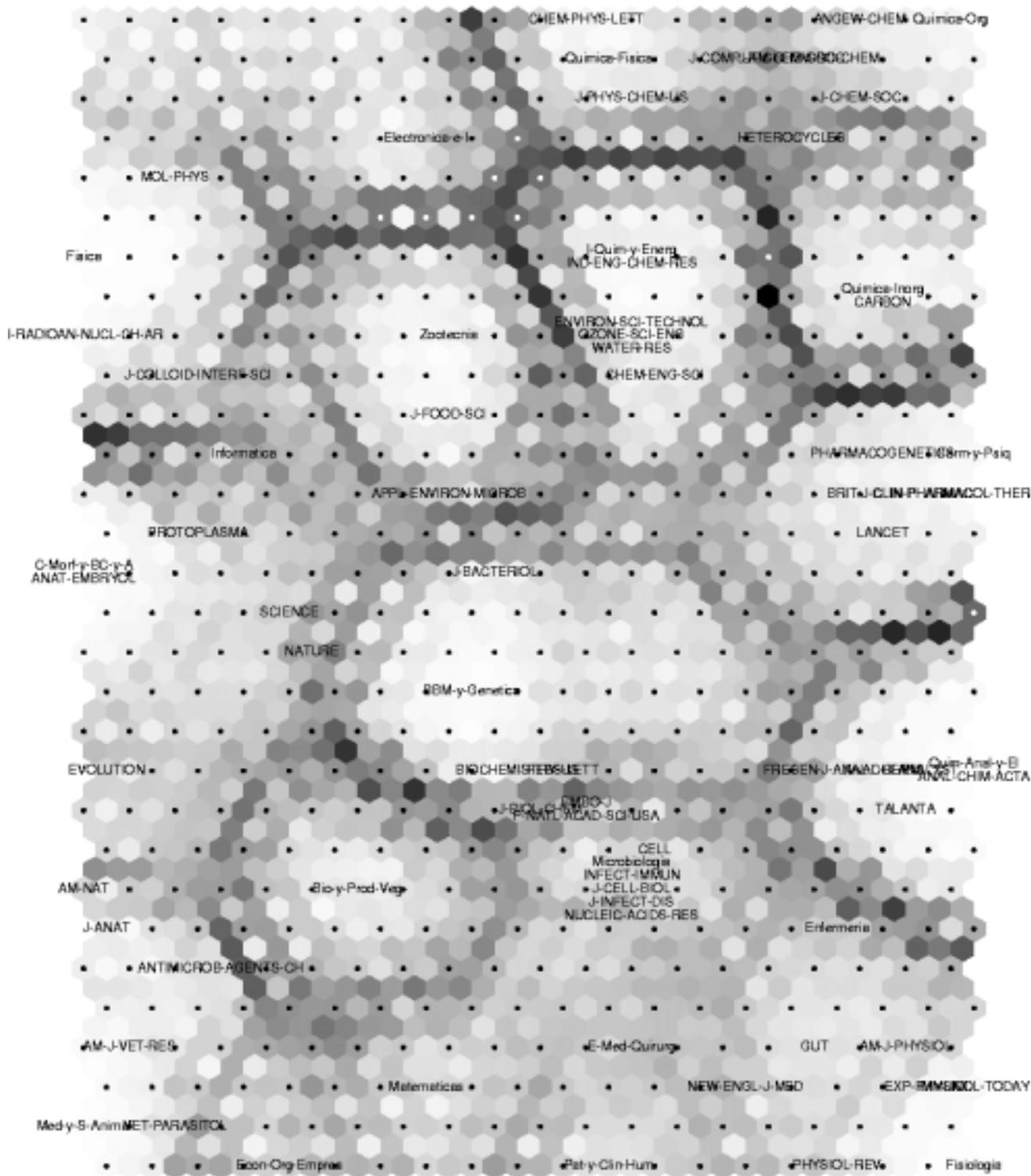


Figura 7