

Programación en R del método de las palabras asociadas

Co-word Method in R

DANIEL HERNANDO RODRÍGUEZ^a
AUTOR

CAMPO ELÍAS PARDO^b
DIRECTOR

DEPARTAMENTO DE ESTADÍSTICA, FACULTAD DE CIENCIAS, UNIVERSIDAD NACIONAL DE COLOMBIA, BOGOTÁ, COLOMBIA

Resumen

Se describe el método de las palabras asociadas (MPA) presentado en Courtial (1990). Se presenta el paquete `mpa` desarrollado en el *software* R que ejecuta el MPA y con la cual se analizan los temas presentes en los artículos publicados en la Revista Colombiana de Estadística. El análisis de estas temáticas se complementa con el análisis de correspondencias simples de las tablas de autores por año y palabras por año.

Palabras clave: *Software* estadístico, cienciometría, análisis estadístico de textos, clasificación jerárquica.

Abstract

Co-word method presented by Courtial (1990) is described. Also the package `mpa` developed in R software is presented, an analysis of topics in Revista Colombiana de Estadística publications is made with this package. The analysis of these subject matters complements itself with simple correspondences analysis of the authors-year and keywords-year tables.

Key words: Statistical software, scientometrics, statistical textual analysis, hierarchical clustering.

1. Introducción

En este documento se presenta el método de las palabras asociadas (MPA) el cual tiene como objetivo principal encontrar las diversas temáticas presentes dentro de un conjunto de documentos, a partir de la clasificación de las palabras clave que se encuentran simultáneamente en éstos. Además permite encontrar la estructura de las relaciones entre los referentes de esas palabras (Charum 1998).

Estas temáticas se hallan con la construcción de grupos de las palabras clave, los cuales se caracterizan en un mapa estratégico para su mejor análisis. Un mapa estratégico es un diagrama de dos ejes en donde se ubican los grupos formados, el eje vertical representa la densidad de los grupos (intensidad de las asociaciones internas) y el eje horizontal representa la centralidad de los mismos (relación de un grupo con los otros), estos conceptos se ampliarán más adelante. El análisis de los grupos se profundiza con el diagrama de la red de relaciones entre las palabras de cada grupo, que muestra a manera de *grafo*, el nivel de asociación entre las palabras clave que pertenecen a cada clase.

De esta manera, el MPA resulta ser un método práctico e indispensable al momento de analizar áreas del conocimiento a partir de sus documentos escritos.

^aEstudiante de estadística. E-mail: dhrodriguez@unal.edu.co

^bProfesor asociado. E-mail: cepardot@unal.edu.co

Uno de los programas de cómputo más utilizados para realizar el MPA es el *Leximappe* (Whittaker 1988), que fue desarrollado por la Ecole des Mines en París, pero este software a pesar de ser una gran herramienta es muy antiguo y otros software más modernos que aplican el MPA no son libres o de fácil acceso. Entonces se desarrolla un código en el software R (R Development Core Team 2007) para ejecutar el método de una forma eficaz y que sirva de herramienta básica para futuros análisis del mismo tipo. Así mismo, se utiliza este código para revelar y analizar las áreas de investigación presentes en las publicaciones de la Revista Colombiana de Estadística.

2. Descripción del método

El MPA es un método de análisis cuantitativo que consiste en la clasificación de palabras clave (o descriptores de documentos escritos) en un conjunto de documentos. Los grupos formados después de dicha clasificación son temáticas que se interpretan como áreas del conocimiento presentes en el grupo de documentos.

El proceso para desarrollar el método es el siguiente, (Charum 1995):

- Se parte de una tabla \mathbf{X} de n documentos por m palabras clave (tabla léxica). Cada entrada de esta tabla es 1 (uno) si el documento asume una palabra clave o 0 (cero) si no.
- Se construye la matriz de co-ocurrencias de las palabras $\mathbf{C}=\mathbf{X}'\mathbf{X}$. El número de co-ocurrencias entre dos palabras, es el número de veces que aparecen las dos palabras juntas en todos los documentos.
- Para cada par de palabras se calcula el índice de asociación E_{ij} dado por:

$$E_{ij} = \frac{c_{ij}^2}{c_i c_j} \quad (1)$$

donde c_{ij}^2 es el cuadrado de la co-ocurrencia entre la palabra i y la palabra j (fila i columna j de \mathbf{X}) y c_i y c_j son las frecuencias absolutas de cada una de las palabras en \mathbf{X} . Esto se puede expresar de forma matricial si se construye la matriz $\mathbf{C}^* = \{c_{ij}^2\}$, entonces la matriz simétrica de índices de asociación \mathbf{E} es:

$$\mathbf{E} = \text{Diag}^{-1}\{c_i\}\mathbf{C}^*\text{Diag}^{-1}\{c_i\} \quad (2)$$

Cada coeficiente de asociación es el producto entre la probabilidad que aparezca la palabra j cuando se presenta la palabra i y la probabilidad de tener la palabra i cuando se presenta la palabra j , por lo cual varía entre cero y uno. Es un índice de similitud entre las palabras clave el cual se puede utilizar para la aplicación de métodos de clasificación, (Charum 1998).

Este índice de similitud muestra que dos palabras clave se encuentran cercanas en la medida en que aparezcan simultáneamente en un gran número de documentos, así, en el MPA se realiza una clasificación jerárquica mediante enlace simple (Lebart et al. 1995), dos palabras se agrupan si son las más cercanas en términos de su asociación.

La matriz \mathbf{E} , puede interpretarse como un *grafo*, donde los nodos son las palabras clave y sus vínculos son las asociaciones entre ellas, el método corta este *grafo* en subconjuntos de palabras relacionadas entre sí.

- Cada grupo se forma teniendo en cuenta su tamaño, es decir, se predetermina un número umbral de palabras que pertenezcan al grupo.

Al final del procedimiento se encontrarán diferentes grupos o *clusters* conformados por las palabras más asociadas entre sí, que reflejarán las temáticas presentes en el corpus.

Según Courtial (1990), éste método presenta una ventaja sobre la clasificación común, y es que no se obliga a reagrupar a las palabras clave en grupos cada vez más homogéneos, por el contrario, debido

al enlace simple, en algunos casos se encuentran grupos muy heterogéneos, pero sin embargo, conservan una función esencial en la lógica de las redes: la de enlazar segmentos de palabras de una forma natural.

En este análisis es más importante analizar grupos que contienen las palabras más cercanas entre sí, debido a que éstas son utilizadas simultáneamente con más frecuencia por autores que hablan de temas relacionados. Si se llegara a intentar una clasificación diferente al enlace simple, se podría caer en eliminar la unión de temas relacionados dentro del campo de trabajo de los autores.

2.1. Concepto de densidad, centralidad y diagrama estratégico

La caracterización de cada uno de los grupos de palabras clave formados con el método, se hace a partir de las relaciones internas de cada grupo y de las relaciones entre grupos, a continuación se definen los conceptos de densidad y centralidad, (Courtial 1990).

2.1.1. Densidad de un grupo

La densidad es una medida de la fuerza de las asociaciones internas de un grupo o *cluster*, se define como el promedio de los coeficientes de asociación entre las palabras clave dentro del grupo. Si S es un grupo creado, entonces su densidad D_S es:

$$D_S = \frac{1}{m'} \sum_{i \in S} \sum_{\substack{j \in S \\ j > i}} E_{ij} \quad (3)$$

donde m' es el número de coeficientes de asociación internos no nulos. De esta forma, si las palabras dentro de un grupo aparecen con alta frecuencia de forma simultánea en diferentes documentos, significa que el grupo está representando a una temática elaborada y tendría una densidad alta. Por otro lado, si las palabras dentro del grupo están presentes de forma simultánea sólo en algunos documentos, pero además se encuentran en otros documentos asociadas con otras palabras, se dice que el grupo representa a una temática poco elaborada y por lo tanto, su densidad es baja.

La densidad es importante en el momento de caracterizar un grupo de palabras, porque refleja si la temática que evidencia el mismo, está desarrollada o no.

2.1.2. Centralidad de un grupo

La centralidad mide el nivel de relación de un grupo con los demás. Se calcula como el valor medio de los coeficientes de asociación entre las palabras clave de un grupo con las palabras clave que pertenecen a los demás grupos existentes. Es decir, como su nombre lo indica, la centralidad muestra la importancia de la temática en general. Si S es un grupo creado, entonces su centralidad C_S es:

$$C_S = \frac{1}{m''} \sum_{i \in S} \sum_{j \notin S} E_{ij} \quad (4)$$

donde m'' es el número de coeficientes de asociación externos no nulos. Si un grupo tiene un índice de centralidad alto, significa que la temática representada por éste tiene un alto impacto sobre las demás temáticas, por otro lado, si sucede lo contrario, la temática es poco central.

2.1.3. Diagrama estratégico

Como un *cluster* se caracteriza con las medidas de densidad y centralidad, es posible plasmar el conjunto de *clusters* en un plano bidimensional donde el eje vertical representa la densidad y el eje horizontal representa la centralidad (Charum 1998).

Es importante mencionar que en el diagrama estratégico, cada grupo queda reducido a un punto en el plano, el cual se puede nombrar con la palabra cuya suma de asociaciones internas sea más alta o con un nombre que exprese la temática presente en el grupo.

En este punto el análisis de los grupos se facilita dependiendo de su ubicación en el diagrama. Por ejemplo, si un grupo se encuentra en la parte superior derecha del diagrama (cuadrante 1), se dice que la temática que representa esta desarrollada y es de alta importancia para las demás (ver figura 1) (Cardona 2001).

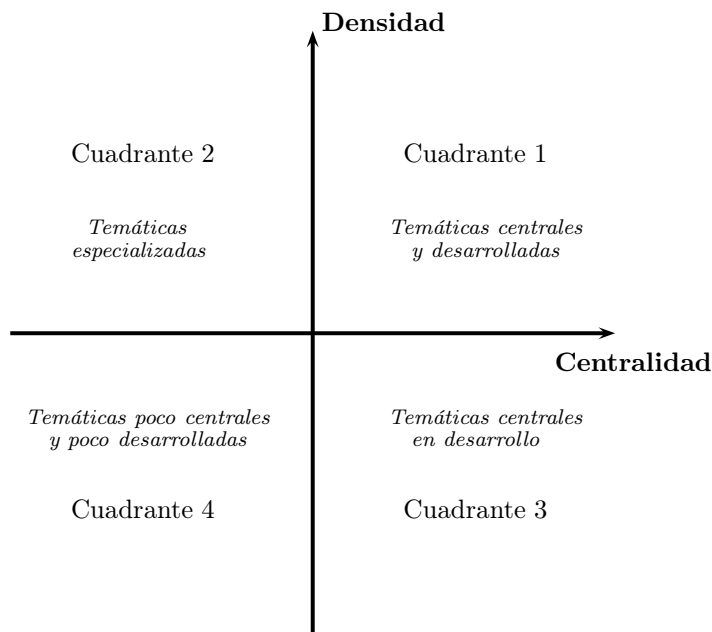


FIGURA 1: Diagrama estratégico

El análisis interno de cada grupo se hace dibujando el *grafo* o red de las asociaciones entre las palabras clave que pertenecen a cada uno de los grupos. Cada vínculo representa la asociación entre cada par de palabras clave, por lo tanto aunque la red de relaciones muestre diferentes uniones, no necesariamente estas uniones son iguales o tienen la misma intensidad para cada par.

3. El método programado en R

El paquete creado en R, se llama `mpa`. Cuenta con cinco funciones básicas que se muestran en la tabla 1. El código principal que realiza la clasificación se encuentra dentro de la función llamada `mpa` la cual requiere la definición de algunos parámetros importantes y del algoritmo base del MPA, los cuales se presentan en esta sección.

La función que grafica la red de un grupo se llama `plotmpa` utilizando la función `plot.network` del paquete `network` de R (Butts et al. 2007). Las demás funciones no requieren de otro paquete especial, fueron programadas con instrucciones básicas del lenguaje propio de R. Algunos pasos de las funciones `leer.mpa` y `matriz.mpa`, se crearon utilizando como referencia una antigua librería para el análisis estadístico de textos llamada `ttta` (Mueller 2004). No se profundiza más en las funciones y no se habla de sus parámetros debido a que existe una documentación de cada una de las funciones del paquete en R, el cual se instala a partir del archivo comprimido (*zip*) que está disponible en la página:

<http://www.docentes.unal.edu.co/cepardot/docs>

TABLA 1: Funciones creadas en R para el MPA y su descripción.

Función en R	Descripción
<code>leer.mpa</code>	Lee las líneas de un archivo de texto en un formato específico.
<code>matriz.mpa</code>	Construye las matrices de co-ocurrencias y asociación, y el diccionario.
<code>mpa</code>	Realiza la clasificación de las palabras clave.
<code>diagram.mpa</code>	Dibuja el diagrama estratégico de los grupos.
<code>plotmpa</code>	Dibuja la red de un grupo creado.

3.1. Parámetros externos del MPA

La clasificación de las palabras clave está determinada por parámetros que controlan que cada grupo pueda ser interpretado de la mejor forma posible. Así, los parámetros externos que se requieren para ejecutar el MPA son (Charum 1998):

1. t_{max} : es el número de palabras máximo por cluster. Si este número es muy grande, se puede hacer difícil el análisis y la interpretación de la temática presente.
2. t_{min} : es el número mínimo de palabras por cluster. Si este número es muy pequeño, la temática que representa no contribuirá mucho al análisis del conocimiento presente en el corpus.
3. f_{min} : es la frecuencia mínima de una palabra clave en el corpus.
4. c_{min} : es el número mínimo de las co-ocurrencias entre dos palabras.

3.2. Algoritmo de clasificación

El algoritmo de clasificación sigue los siguientes pasos:

1. Construcción de la tabla de documentos por palabras clave \mathbf{X} .
2. Eliminación de palabras clave dentro del corpus que tengan frecuencia menor a f_{min} .
3. Construcción de la matriz de co-ocurrencias \mathbf{C} , a partir del corpus.
4. Eliminación de las co-ocurrencias menores a c_{min} dentro de \mathbf{C} .
5. Construcción de la matriz de asociaciones \mathbf{E} , a partir de la matriz \mathbf{C} .
6. Organización decreciente de los $m(m-1)/2$ coeficientes de asociación diferentes.
7. La primera pareja en ser agrupada en una clase, es la que tiene mayor coeficiente de asociación, las demás palabras se van agrupando en las clases siguiendo el orden de las asociaciones hecho en el paso 6.
8. Cuando un grupo llega a tener t_{max} número de palabras clave, no se permite la entrada de una nueva palabra clave dentro de éste.
9. El algoritmo se detiene cuando se hayan evaluado todas las asociaciones no nulas del orden creado en el paso 6.

Debido a la naturaleza del algoritmo anterior, se presenta unas dificultades una vez hecha la clasificación. Como el criterio de creación de cada grupo es un tamaño predeterminado, no se puede garantizar que todas las palabras clave queden clasificadas en algún *cluster*, pueden existir palabras que solamente se encuentren relacionadas con otras que ya pertenezcan a un grupo cuyo tamaño es t_{max} (el umbral). Por lo tanto, este tipo de palabras no van a resultar clasificadas al final del proceso.

Estas palabras clave se llamarán *huérfanas*, la cantidad de huérfanas en un proceso de clasificación depende de la estructura de red que presenten las palabras clave y de la buena escogencia del parámetro t_{max} .

4. Análisis de los temas presentes en las publicaciones de la Revista Colombiana de Estadística

La Revista Colombiana de Estadística es relativamente joven si se compara con otras publicaciones de carácter científico. Lanzó su primer ejemplar en el año 1968, el cual constaba de tan sólo dos artículos, el segundo número salió en el año siguiente y sólo tenía un artículo. Desde esa fecha la Revista estuvo inactiva y hasta 1978 salió una nueva publicación. Sólo se logró una continuidad en las publicaciones hasta el año 1980, y desde entonces se publican dos números cada año.

En total, se tienen 244 artículos publicados desde 1968 hasta el primer número de 2007. El corpus contiene un total de 816 palabras clave diferentes, de las cuales 679 son hápax, es decir, aparecen una sola vez en todo el corpus, 137 palabras aparecen por lo menos 2 veces y 48 palabras aparecen por lo menos 3 veces.

En la tabla 3 se presenta un resumen del número de palabras clave según su frecuencia. En el corpus inicial, la palabra que más se repite tiene frecuencia 15 y corresponde a la palabra clave *teaching statistics*. En esta tabla se evidencia una gran dispersión de palabras clave dentro del corpus, muchas palabras distintas y pocas que se repiten, esto también puede deberse al tamaño tan pequeño del corpus en estudio y a la diversidad de áreas en las que actúa la estadística.

4.1. Criterios para el análisis

En Whittaker et al. (1989), se discute sobre dos formas de abordar un análisis utilizando el método de palabras asociadas. El primero, es realizar el análisis sobre los descriptores o palabras clave de los documentos y el segundo es hacerlo utilizando las palabras de los títulos de los documentos.

Las dos formas de análisis tienen sus ventajas y desventajas. En el primero, se presenta una desventaja si para un conjunto de documentos, un grupo de *indexadores* revisa cada documento y le asignan a su vez los descriptores según su criterio. Al descubrir y analizar las temáticas con el MPA, puede caerse en un posible sesgo sistemático: revelar la conceptualización de los indexadores sobre los documentos y no el verdadero conocimiento presente en éstos. Whittaker et al. (1989) lo llaman el *efecto del indexador*, y aunque no se tiene control sobre este efecto, se debe partir de la premisa de que los *indexadores* eligen cuidadosamente las palabras clave que asignan, y éstas son de hecho, indicación fiable de los conceptos científicos a los que se hace referencia en los documentos.

En la Revista Colombiana de Estadística, los primeros artículos publicados no tienen palabras clave debido a su antigüedad. Las palabras clave que aparecen en el corpus fueron asignadas por un grupo de profesores del Departamento de Estadística. En este caso, hay que partir de la premisa mencionada en el párrafo anterior.

La segunda manera de análisis presenta un problema un poco más grave: las palabras del título no son estándar, en los títulos se pueden utilizar diferentes palabras que hablan de un mismo tema, lo que disminuye la homogeneidad en el corpus. Además, debido a que el criterio de formación de un *cluster* es el número de palabras que pertenecen a éste, se tendría que ampliar este parámetro ya que el número de palabras aumenta, y las clases resultarían difíciles de interpretar.

Para el caso de la Revista, se aborda una metodología conjunta a las dos mencionadas. Los autores de cada artículo de la Revista proponen sus propias palabras clave, pero la Revista sigue las reglas del Current Index to Statistics¹, y en una de sus reglas dice que las palabras clave deben ser diferentes a las

¹<http://www.statindex.org/>

palabras del título. En este caso, posibles palabras que describen el tema en el título, no se encuentran contempladas dentro del corpus inicial.

Para solucionar este problema, se realizó un minucioso trabajo de adicionar a las palabras clave ya existentes, palabras del título que indiquen el tema tratado en el artículo y que no halla sido contemplado en el corpus. Además de una labor de homogeneización debido a que las palabras clave utilizadas en la Revista no son estándar, existen dentro del corpus inicial, palabras clave que reflejaban en esencia el mismo tema pero son gramaticalmente diferentes, incluyendo palabras que en ocasiones se utilizan en singular y otras en plural.

Después de esta homogeneización del corpus, resultó un total de 776 palabras clave diferentes, de las cuales 638 son hápax, 138 palabras aparecen por lo menos 2 veces y 59 palabras aparecen por lo menos 3 veces (ver tabla 3). Para este nuevo corpus, la palabra más frecuente es *time series* la cual aparece 17 veces.

Luego de ver la tabla, se evidencia que aún persiste una gran frecuencia de palabras diferentes, y una baja frecuencia de palabras que se repiten, lo cual hace que el análisis sea débil si se llegara a aplicar al corpus tal como está. Por lo tanto se decidió hacer una nueva adición de palabras clave, se asignó a cada documento una palabra clave que reflejara un tema general dentro de la estadística. Para esto, se utilizó como referencia la clasificación de palabras clave de la MSC2000 (Mathematics Subject Classification 2000)², y de esta forma integrar a cada documento dentro de un tema general, las palabras introducidas en el corpus se muestran la tabla 2.

TABLA 2: Palabras generales introducidas al corpus.

Código	Palabra MSC2000	Palabra introducida
62A01	Foundational and philosophical topics, Historical	Statistical foundations
62D05	Sampling theory, sample surveys	Sampling theory
62Exx	Distribution theory	Distribution theory
60-xx	Probability theory	Probability theory
62Fxx	Parametric inference	Statistical inference
62Gxx	Nonparametric inference	Nonparametric statistics
62Hxx	Multivariate analysis	Multivariate analysis
62Jxx	Linear inference, regression	Regression analysis
62Kxx	Design of experiments	Design of experiments
62Mxx	Inference from stochastic processes	Stochastic processes
62M10	Time series	Time series
62Nxx	Survival analysis	Survival analysis
62Pxx	Applications	Applications

Una vez hecha esta última agregación, el corpus final contiene 774 palabras diferentes, de las cuales 638 son hápax, 136 aparecen por lo menos dos veces y 63 aparecen por lo menos 3 veces (ver tabla 3). Las palabras más frecuentes son *statistical inference* que aparece 35 veces, *regression analysis* 34, *applications* 32, *statistics foundations* 28, *stochastic processes* 22, *design of experiments* 20, *time series* 20, *multivariate analysis* 18 y *teaching statistics* 15.

Para el análisis de las publicaciones de la Revista Colombiana de Estadística, los parámetros del MPA se eligieron así:

$$t_{max}=10, f_{min}=3, c_{min}=1, t_{min}=3.$$

Las instrucciones básicas para realizar la clasificación con el paquete *mpa*, se muestran en el apéndice.

4.2. Resultados

En esta sección se muestran los resultados del MPA en las publicaciones de la Revista, así como los resultados del análisis de correspondencias simples hecho para las tablas de palabras contra años y autores

²<http://www.ams.org/msc/>

TABLA 3: Número de palabras clave y su frecuencia dentro del corpus.

Corpus inicial		Corpus luego de la lematización		Corpus final	
No. de palabras clave	Frecuencia	No. de palabras clave	Frecuencia	No. de palabras clave	Frecuencia
679	1	638	1	638	1
89	2	79	2	73	2
25	3	32	3	27	3
6	4	6	4	5	4
4	5	4	5	4	5
1	6	4	6	8	6
3	7	3	7	2	7
1	8	1	8	2	8
1	9	1	9	2	10
2	10	2	10	1	11
2	11	1	11	1	12
2	12	2	12	2	14
1	15	1	14	1	15
-	-	1	15	1	18
-	-	1	17	2	20
-	-	-	-	1	22
-	-	-	-	1	28
-	-	-	-	1	32
-	-	-	-	1	34
-	-	-	-	1	35

contra años, que se realizó para enriquecer el análisis.

Dados los parámetros definidos en la sección anterior, para el algoritmo de clasificación entraron un total de 63 palabras clave distintas. Se obtuvo un total de 7 clases, con un promedio de 8.7 palabras por clase. La densidad promedio de los grupos creados fue de 0.081 y la centralidad promedio de 0.021. En la tabla 4 se muestran los resultados de la clasificación, los grupos presentados están ordenados tal y como se formaron dentro del algoritmo. El número de huérfanas en la clasificación es 2, estas palabras son *linear programming* y *methodology*, ambas con frecuencia 3.

TABLA 4: Resumen de las temáticas de investigación dentro de la Revista Colombiana de Estadística.

Cluster	Tamaño	Densidad	Centralidad
Statistics foundations	10	0.126	0.019
Design of experiments	6	0.065	0.023
Least squares method	10	0.099	0.028
Statistical inference	10	0.071	0.022
Probability theory	5	0.075	0.018
Stochastic processes	10	0.080	0.019
Sampling	10	0.047	0.017

La figura 2 muestra el diagrama estratégico, el cual describe cómo se posicionan las diferentes temáticas encontradas. En el cuadrante 1 se ubica tan sólo un grupo, el cual es el más central que el resto: *Least Squares Method*. Se muestra como una temática madura y que tiene más alta relación con las demás debido a que su densidad y centralidad son más altas que el promedio.

En el cuadrante 2 se ubica la temática *Statistics Foundations*. Es un campo de investigación especializado, una temática muy publicada pero que permanece relativamente aislada del resto. En el cuadrante 3, aparecen las temáticas llamadas *Statistical Inference* y *Design of Experiments*. Estas temáticas son de gran importancia general pero no han sido muy trabajadas, se puede pensar que en su desarrollo cautiven a otros autores y pasen a ubicarse en el cuadrante 1.

Por último, en el cuadrante 4 se ubican las temáticas menos centrales y menos densas: *Stochastic Processes*, *Probability Theory* y *Sampling*. Son campos que no han sido muy trabajados dentro de la Revista y tampoco tienen importancia relativa frente a los demás, estas temáticas pueden extenderse en el futuro hacia alguno de los otros cuadrantes o simplemente desaparecer.

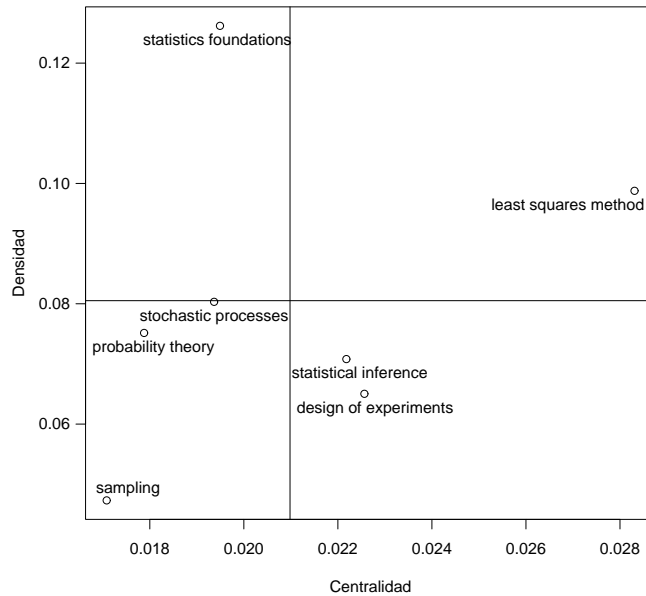


FIGURA 2: Diagrama estratégico de las áreas de investigación presentes en la Revista.

En general, los resultados manifiestan que aún la Revista se encuentra en un proceso de crecimiento. Aunque uno de los grupos se muestre central y otro denso en el diagrama, se observa que en realidad la centralidad y la densidad de las temáticas en general son bajas. El grupo *Least Square Method* que tiene la centralidad más alta, tan sólo alcanza el valor 0.028, lo cual muestra que en realidad hay una ausencia de temáticas centrales dentro de la Revista. Algo similar sucede con el grupo *Statistics Foundations*, su densidad es la más alta pero es sólo 0.126. Es la más especializada del resto, pero es aventurado decir que es una temática realmente especializada.

En la siguiente sección se analizará con más profundidad cada una de las temáticas encontradas, a fin de revelar la estructura interna de los grupos y la intensidad de sus vínculos.

4.3. Temáticas dentro de la Revista: ¿una motivación a publicar?

La importancia de los hallazgos de este trabajo recae directamente sobre los autores de la Revista, además es una motivación a los investigadores que trabajan en las áreas mostradas pero que no publican en éstas. Una vez visto el diagrama estratégico, se descubre que muchos de los campos de la estadística no han sido publicados: aparecen en el cuarto cuadrante del diagrama, o sencillamente no se descubren en el análisis.

Al establecer los temas implícitos dentro de cada grupo, se esclarecerá el área de la estadística que se está posicionando dentro del diagrama estratégico y es una alerta (o todo lo contrario) para los investigadores de cada área en cuestión.

En esta sección, cada gráfico muestra la red de relaciones de los temas dentro de cada grupo, en color rojo (gris en blanco y negro) la palabra clave que le da el nombre, es decir, la palabra cuya suma de asociaciones internas es mayor, lo cual no significa que esta palabra refleje totalmente la temática presente en el grupo. Los nodos de las redes son las palabras clave y los segmentos que los unen indican si hay o no relación entre éstos, además el grosor de cada segmento evidencia si la asociación es fuerte o débil.

4.3.1. Temáticas maduras y centrales

En la figura 3 se muestra la red del grupo *Least Squares Method*. La estructura interna refleja claramente el campo de investigación que se revela en esta clase: el análisis de regresión. Las palabras entorno a *regression analysis* muestran la claridad de la interacción entre los temas presentes, se manifiesta el campo trabajado desde tres puntos de vista: primero la forma de estimar un modelo de regresión (*least squares method, least squares estimator*), segundo las pruebas de hipótesis (*hypothesis testing*) y finalmente el ajuste del modelo de regresión (*goodness of fit, outliers, dfbeta statistic, influential observations*). La centralidad del grupo demuestra que el campo no solamente es publicado con los temas presentes sino que existe una interacción con los demás temas de la Revista.

En el grupo también aparece un tema al parecer intruso dentro del campo: *nonparametric statistics*, es un tema que sólo está relacionado con la prueba de hipótesis y por esto se encuentra en este grupo. El hecho de que *nonparametric statistics* no se encuentre clasificado en otro *cluster*, donde interactúe con temas más afines, demuestra claramente que dentro de la Revista no ha sido un tema muy publicado, siendo así, que sólo se presenta como un tema desde el punto de vista del juzgamiento de hipótesis.

4.3.2. Temáticas especializadas

La red de las relaciones internas de la temática *Statistics Foundations* se muestra en la figura 4. *Statistics Foundations* es el tema con densidad más alta que el resto, lo que sugiere que dentro de la Revista ha sido más trabajada y profundizada, pero el campo como tal no tiene una importancia relativa frente al resto. Se representa una de las grandes áreas de la Revista, en donde se publica el estado de la estadística en cuanto a la evolución en la carrera, en la educación, en la forma de enseñar estadística, además de la recopilación de los trabajos de los estadísticos colombianos y extranjeros. Es por esto que es un área en continuo desarrollo, las publicaciones dentro de esta temática se presentan tal y como va evolucionando la estadística misma, desde el punto de vista de la educación y sus profesionales. La red muestra la dinámica e intensidad de las relaciones entre estos temas.

4.3.3. Temáticas centrales en desarrollo

En la figura 5 se muestran las redes de las temáticas centrales. La clase *Statistical Inference* muestra una de las áreas más importantes dentro de la estadística, la inferencia estadística a pesar de ser un campo muy importante, no ha sido muy trabajada dentro de la Revista, sin embargo, debido a la naturaleza del conocimiento presente en este grupo y de la centralidad del mismo demuestran que es un tema de importancia general para las publicaciones analizadas.

La estructura de red muestra la dinámica de la inferencia interactuando desde diferentes puntos de vista, se revelan temas como la estimación, la construcción de intervalos de confianza y los estimadores máximo-verosímiles. Temas como la teoría de las distribuciones, la simulación y los métodos estadísticos matizan el campo de acción de la inferencia dentro de la Revista.

El grupo *Design of Experiments* es un tema muy relacionado con *Least Square Method*. A pesar de no tener muchas palabras clave dentro de sí, se muestra el núcleo del tema trabajado, el diseño de experimentos interactuando directamente con los modelos lineales, el análisis de varianza y el concepto de varianza, el cual es muy general y tal vez debido a eso y al otro tema general, la variabilidad, esta temática se posiciona como central. Aunque dentro de la Revista existen muchos temas que pueden interactuar con el diseño de experimentos y enriquecer el conocimiento, es muy poco lo que se ha trabajado y por eso es que este grupo no es tan abundante de temas como otros.

4.3.4. Temáticas poco centrales y poco desarrolladas

La figura 6 muestra las temáticas de esta sección. El grupo *Stochastic Processes* se mantiene en el promedio de la densidad, mostrando que es un grupo que tiende a ser maduro. Los vínculos internos reflejan la forma particular en la que se creó este grupo, la interacción de dos de las área más importantes

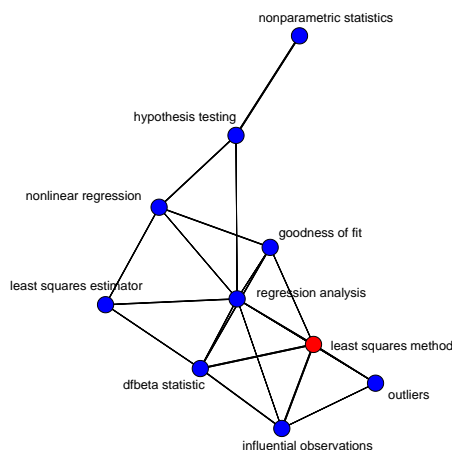


FIGURA 3: Red de la clase Least Squares Method.

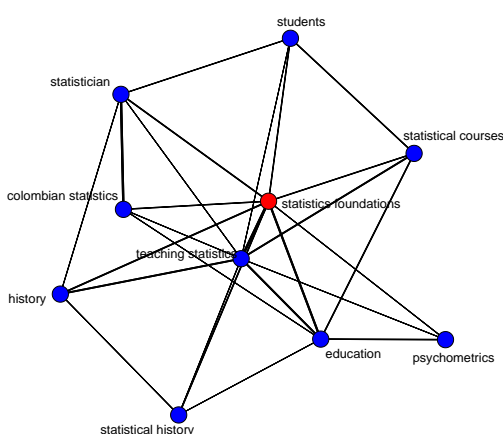


FIGURA 4: Red la clase Statistics Foundations.

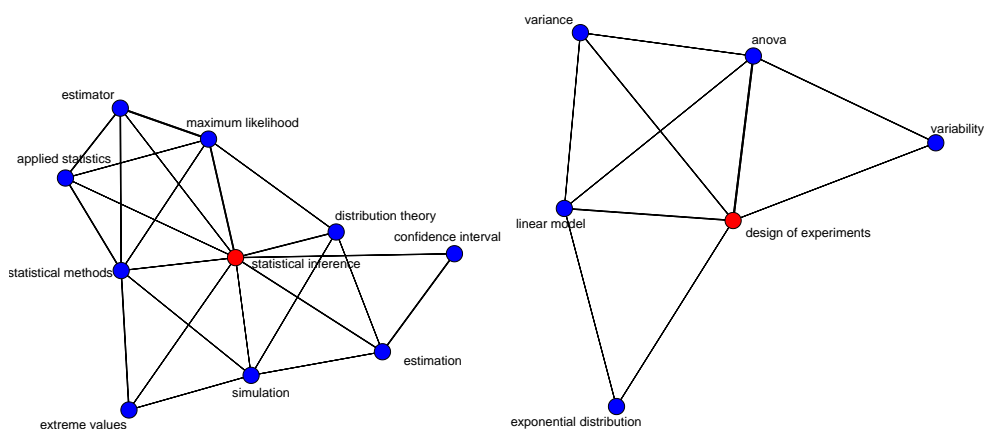


FIGURA 5: Redes de la clases Statistical Inference y Design of Experiments.

de la estadística: los procesos estocásticos y el análisis de series de tiempo. Estos dos temas se muestran estrechamente asociados dentro del grupo, al parecer, dentro de la Revista los procesos estocásticos han sido más enfocados en las series de tiempo, donde interactúa con temas como pronóstico, modelos estocásticos y modelos ARIMA. También sobresalen temas como las cadenas de markov y martingalas.

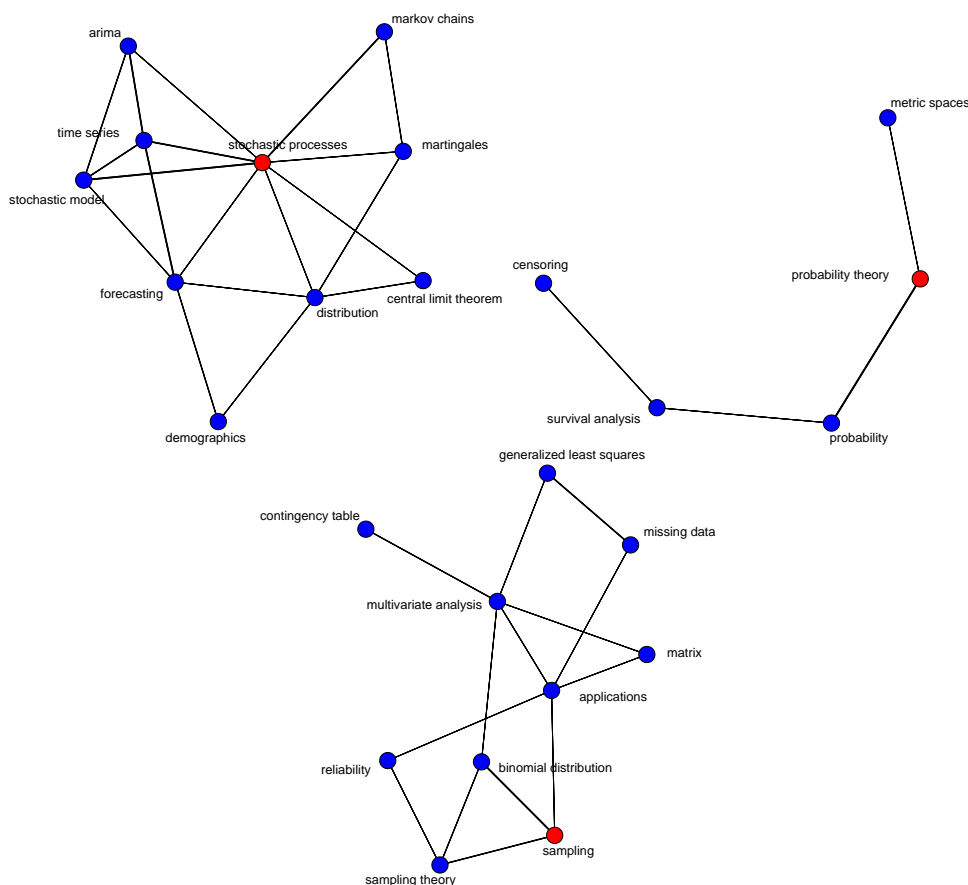


FIGURA 6: Redes de la clases Stochastic Processes, Probability Theory y Sampling.

En este grupo aparece el tema *demographics*, el cual sólo está relacionado con pronóstico y distribución. El hecho de que aparezca esta palabra sugiere que ha sido investigado dentro de la Revista, pero no dentro desde el punto de vista de los procesos estocásticos, tema con el cual no se mantiene ninguna relación dentro de la Revista.

El grupo *Probability Theory*, se muestra con una característica particular: la forma de cadena sugiere que no existe una real interacción entre los temas presentes, cada uno está relacionado con otros dos lo que muestra que los temas que no están enlazados pueden ser potencialmente investigados en futuros trabajos.

Finalmente, el grupo *Sampling*, muestra dos grandes áreas de la estadística desde el punto de vista de las aplicaciones, pero que no han sido desarrollados ni son de importancia general dentro de las publicaciones de la Revista. En primera instancia se muestra el análisis multivariado y su aplicación, interactuando con temas como las tablas de contingencia lo que sugiere que está presente de manera implícita el análisis de correspondencias. En segundo lugar se muestra la teoría del muestreo, relacionado con temas como la distribución binomial.

4.3.5. Análisis de correspondencias simples dentro del análisis de las temáticas de la Revista

Se realizó un análisis de correspondencias simples (ACS) para la tabla de contingencia de palabras contra años y luego para la tabla de autores contra años. Debido a la gran dispersión de los datos, se decidió agrupar los años en seis intervalos de tiempo, que van desde 1968 a 1982, 1983 a 1987, 1988 a 1992, 1993 a 1997, 1998 a 2002 y 2003 a 2007. El análisis se llevó a cabo con el paquete `ade4` (Chessel et al. 2004)

y los gráficos de los planos con la función `planfac` del paquete `FactoClass` (Pardo & Del-Campo 2007).

En la figura 7 se muestra el primer plano factorial del ACS para la tabla de palabras contra años, los dos primeros ejes recogen un total de 53.7% de la inercia total. Este análisis provee una ubicación temporal de las temáticas encontradas dentro de la Revista. Se muestra que las palabras pertenecientes al grupo *Statistics Foundations* se encuentran más relacionadas con los primeros años de la Revista, lo cual explicaría su densidad más alta que las demás temáticas. Por otro lado, las palabras relacionadas con la temática *Least Squares Method*, aparecen relacionadas con los años de 1988 hasta 1997, el tema *regression analysis* aparece muy relacionado con el intervalo 1998 a 2002, lo cual explica la centralidad de la temática, pues ha sido trabajada a lo largo del tiempo de la Revista y esto favorece la interacción con otros temas.

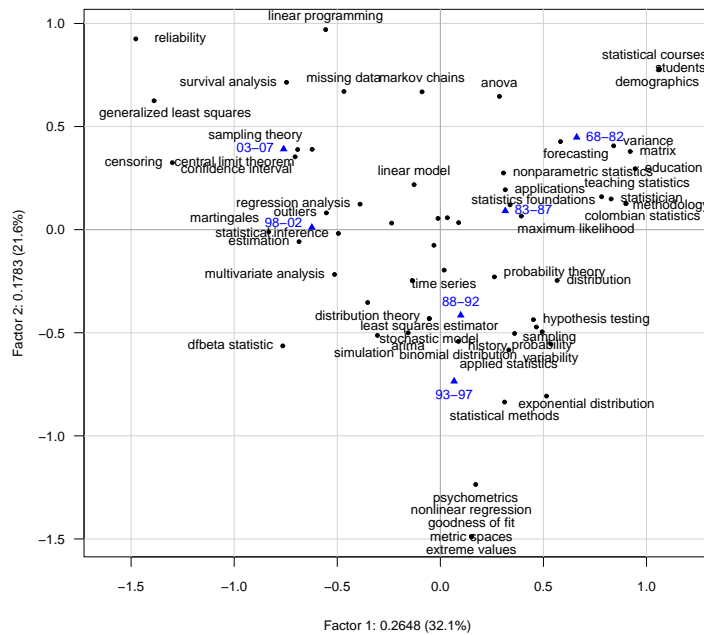


FIGURA 7: Primer plano factorial del ACS para la tabla Palabras-Años.

Al parecer, la temática *Statistical inference* aparece dentro de la Revista en los últimos años, sin embargo, la palabra *statistical methods* aparece unos años antes. Por otro lado, los temas de *Design of Experiments* se muestran en el plano en medio de los años, lo que indica que han sido trabajados a lo largo del tiempo, lo que explica la razón por la cual éste sea un tema central. Algo parecido sucede con la palabra *markov chains*, que pertenece a la temática *Stochastic Processes*, los temas *time series* y *stochastic model* se muestran más asociados a los años 1988-1992. Los temas asociados a los grupos *Probability Theory* y *Sampling*, se encuentran repartidos a lo largo del plano, lo que indicaría que estos temas han sido tratados de una forma esporádica, debido a que son los menos trabajados y desarrollados dentro de la Revista,

Para el ACS para la tabla de autores contra años, sólo entraron al análisis los autores que aparecieran dentro de la Revista más de dos veces. En la tabla 5 se muestran los 32 autores que se tuvieron en cuenta para el análisis y su respectiva frecuencia. En la figura 8 se muestra el primer plano factorial, los dos primeros ejes recogen un total de 65.6% de la inercia total.

El plano muestra el paso de los autores en la Revista Colombiana de Estadística. Se observa que los primeros autores de la Revista son Castillo, Fandiño, Moreno, Cepeda F., y Cortés. Los autores que publicaron en la Revista en los años intermedios son Castaño, Mayorga, Beltrán, Nieto, Pacheco, Corzo,

TABLA 5: Autores más frecuentes dentro de la Revista Colombiana de Estadística hasta el primer número del año 2007.

Autor	Frecuencia
López L. A.	14
Vargas J.A.	12
Ospina D.	10
Nieto F.	9
Bautista L.	8
Cepeda F.	8
Muñoz M.	8
Ortiz J.	8
Pacheco P. N.	8
Blanco L.	7
Correa J. C.	7
Yañez S.	7
Cortés L.	5
Jiménez J. A.	5
Knolle H.	5
Moreno L.	5
Beltrán G.	4
Chavez B.	4
Corzo J.	4
Fandiño R.	4
Matos R.	4
Melo O.	4
Castañeda J.	3
Castaño E.	3
Castillo M. F.	3
Madan K. C.	3
Marínez J.	3
Mayorga H.	3
Pérez A.	3
Quintero C.	3
Rodríguez I.	3
Velasco A.	3

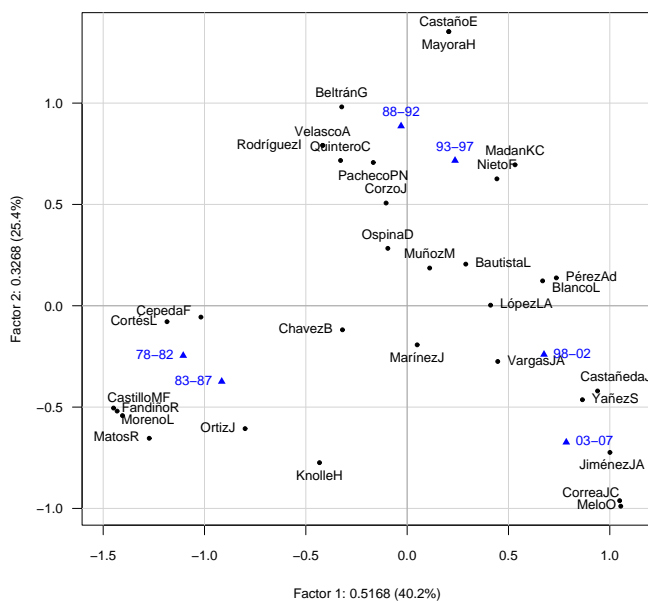


FIGURA 8: Primer plano factorial del ACS para la tabla Autores-Años.

Quintero, Velasco, entre otros. Se puede decir que los autores de los años 1998-2002 son Vargas, Castañeda y Yañes. Finalmente los últimos autores más frecuentes han sido Jiménez, Correa y Melo. Los autores que se encuentran en el medio del plano no se ven claramente a quienes están asociados, al parecer han estado presentes a lo largo de la revista.

5. Conclusiones y comentarios finales

El algoritmo creado en R para el análisis de palabras asociadas, funcionó como una gran herramienta para encontrar los resultados mostrados anteriormente. El paquete `mpa` realiza la clasificación de una forma eficaz, la representación de los grupos en el diagrama estratégico y la forma de mostrar cada una de las redes, hace más fácil el análisis de los temas que se descubren. En el futuro, se puede perfeccionar la herramienta para solucionar el problema de las palabras clave *huérfanas*, además de optimizar algunos procesos internos del código e implementar nuevas metodologías para presentar los resultados.

Los resultados de la aplicación del MPA, reflejan que en realidad la Revista Colombiana de Estadística aún está en crecimiento. Los temas encontrados dentro de la Revista tienen centralidad y densidad bajas. El tema más especializado, tiene que ver con la historia y evolución de la estadística, como carrera y como ciencia. La inferencia estadística y el diseño de experimentos se revelan como los más centrales dentro de la Revista, estos temas tienen gran importancia general y pueden seguir desarrollándose en el futuro. Finalmente, los temas de procesos estocásticos, probabilidad, muestreo y análisis multivariado se manifiestan como las menos trabajadas y desarrolladas en las publicaciones.

En general, las temáticas se encuentran repartidas a lo largo de los años de la Revista, sólo la temática *Statistics Foundations* aparece más asociada a los primeros años, *Statistical Inference* y *Stochastic Processes* aparecen partir del intervalo 1988-1992. El análisis de correspondencias de la tabla Autores-Años, evidencia que en general los autores han pasado por la Revista sin permanecer vigentes en ésta, sólo de muy pocos se puede decir que han estado presentes a través de todos los años, como Ospina D., Muñoz M. y Martínez J., los cuales se posicionan en el centro del plano.

Referencias

- Butts, C. T., with help from David Hunter & Handcock, M. S. (2007), *network: Classes for Relational Data*. R package version 1.2.
- Cardona, P. A. (2001), Comparación entre el análisis estadístico de datos textuales y el método de palabras asociadas, Trabajo de grado para optar al título de estadístico, Universidad Nacional de Colombia. Facultad de Ciencias. Departamento de Estadística, Bogotá.
- Charum, J. y Meyer, J. (1998), *Hacer ciencia en un mundo globalizado. La diáspora científica colombiana en perspectiva*, TM editores en coedición con Colciencias y la Facultad de Ciencias de la Universidad Nacional de Colombia, Bogotá.
- Charum, J. y Parrado, L. (1995), *Entre el productor y el usuario. La construcción social de la utilidad de la investigación*, Instituto Colombiano para el Fomento de la Educación Superior ICFES, Universidad Nacional de Colombia, Bogotá.
- Chessel, D., Dufour, A.-B. & Thioulouse, J. (2004), 'The ade4 package-I- One-table methods', *R News* 4, 5-10.
- Courtial, J. (1990), *Introduction à la scientométrie*, Anthropos - Económica, París.
- Lebart, L., Morineau, A. & Piron, M. (1995), *Statistique exploratoire multidimensionnelle*, Dunod, Paris, p. 156.

Mueller, J.-P. (2004), *ttda: Tools for textual data analysis*. R package version 0.1.1, [citado el 20 de diciembre de 2007].

*<http://wwwpeople.unil.ch/jean-pierre.mueller>

Pardo, C. & Del-Campo, P. (2007), 'Combinación de métodos factoriales y de análisis de conglomerados en R: el paquete **FactoClass**', *Revista Colombiana de Estadística* **30**(2), 231–245.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

*<http://www.R-project.org>

Whittaker, J. (1988), 'Leximape - doc'. Ecole des Mines, París.

Whittaker, J., Courtial, J. P. & Law, J. (1989), 'Creativity and conformity in science: Titles, keywords and co-word analysis', *Social Studies of Science* **19**(3).

Apéndice

```
library(mpa)
#Lectura de los datos-----
datos <- leer.mpa("revista.txt")
#creación de las matriz de asociaciones y co-ocurrencias-----
matriz <- matriz.mpa(datos, fmin=3, cmin=1)
#palabras y su frecuencia-----
d <- diag(matriz$Matrizc)
#clasificación-----
clas <- mpa(matriz$Matriza,10,matriz$Palabras)
#diagrama estratégico-----
diagram.mpa(clas,tmin=3)
#red de cada una de las clases-----
for(i in 1:7){
windows()
plotmpa(i,matriz$Matriza,clas)
}
```