EMPLEO DE TÉCNICAS DE CLUSTERING Y MDS EN LA CARACTERIZACIÓN DE LÓS TEMAS DE INVESTIGACIÓN DE LA PRODUCCIÓN FORESTAL ESPAÑOLA ISI 1994-2003.

J. Pino-Díaz

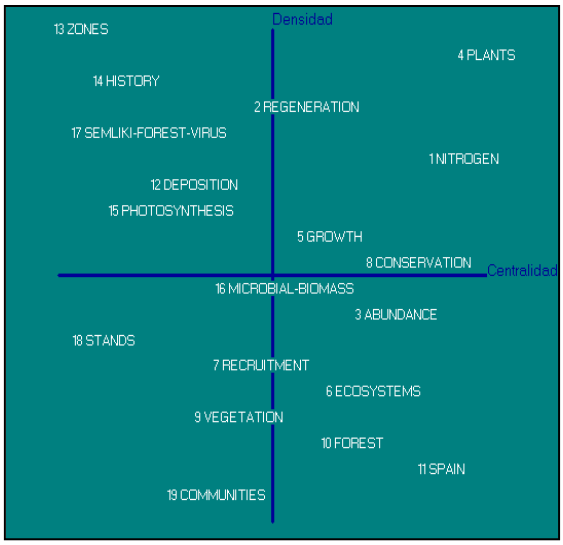
Facultad de Biblioteconomía y Documentación, Universidad de Granada.
Colegio Máximo de Cartuja, 18071-GRANADA
josepinodiaz@wanadoo.es

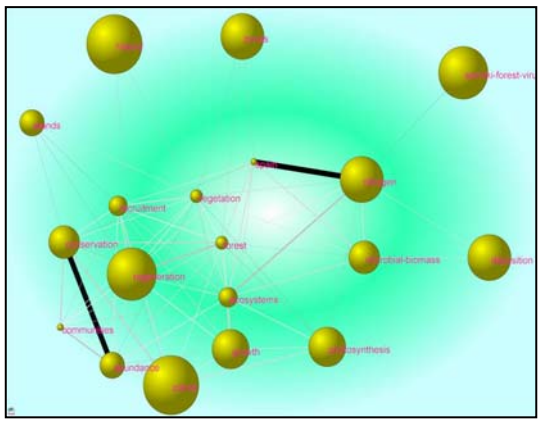
Resumen. En el presente trabajo se realiza un estudio de las relaciones existentes entre los temas de la investigación forestal española del periodo 1994-2003 indexada en ISI. Para ello se han utilizado técnicas de agrupamiento y de escalamiento multidimensional, en las que las variables han sido los temas de investigación y los casos, los enlaces externos entre temas (co-ocurrencia de descriptores), la élite de revistas de publicación o la élite de autores. Según los casos, se han obtenido vinculaciones significativas no coincidentes en los tres tipos de agrupamiento realizado (semántico, de publicación o de autor), salvo en los grupos *nitrogen-spain, abundance-conservation-communities y forest-regeneration-recruitment*. Se observa que un completo análisis de las relaciones entre temas requiere del estudio conjunto de las proximidades semántica, de publicación y de autoría.

Palabras Clave. VIRIs, interfaces gráficos para la visualización de información, clasificación, agrupamiento, escalamiento multidimensional, dendrogramas, gráficos MDS, sociogramas, análisis de palabras asociadas, análisis de redes sociales.

INTRODUCCIÓN

Dos trabajos anteriores sobre la producción científica española ISI 1994-2003 han permitido conocer los diecinueve temas de la investigación forestal española, la red que conforman, su diagrama estratégico y su sociograma, empleando técnicas de análisis de palabras asociadas y de redes sociales.





Otras técnicas estadísticas permiten indagar sobre las relaciones, a veces ocultas, existentes entre datos. Habitualmente se emplean técnicas de clasificación (clustering) o de escalamiento multidimensional (MDS) para conocer y caracterizar estas relaciones. En ambos casos se obtienen gráficos, dendrogramas y gráficos MDS, que facilitan la comprensión de los resultados obtenidos y que pueden ser utilizados como interfaces gráficos para la visualización de la información.

OBJETO

El objeto de este trabajo es visualizar e interpretar las relaciones, a veces "ocultas", existentes entre los diecinueve temas de investigación forestal española del periodo 1994-2003. obtenidos a partir del análisis de palabras asociadas al que se sometió la colección de artículos indexados en la base de datos Science Citation Index. (ISI).

MATERIAL Y MÉTODOS

Se parte de la información previa disponible resultado del análisis de palabras asociadas al que se ha sometido la colección de artículos con origen en España de la base de datos Science Citation Index, del periodo 1994-2003, que tratan el tema *"forest*"*.

Se dispone de la información que caracteriza los 19 temas de investigación. En este trabajo se utilizarán los datos sobre los enlaces externos entre temas de investigación (co-ocurrencia de descriptores), sobre la élite de revistas y sobre la élite de autores de cada tema, para el procesamiento estadístico con los programas STATÍSTICA 6.0 y SPSS 12.0.

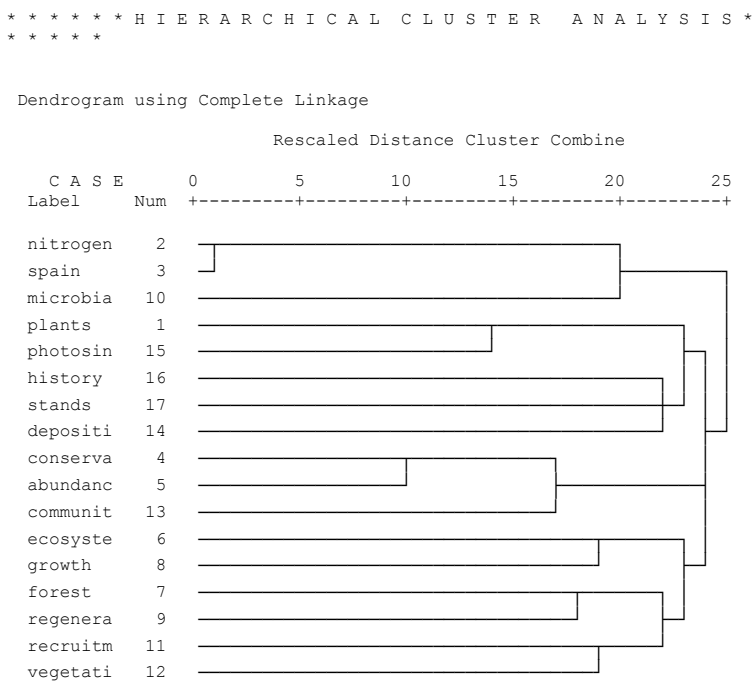
Con el procesamiento estadístico de los datos se han conseguido gráficos (dendrogramas y gráficos MDS) que han permitido observar los agrupamientos de los temas en función de los enlaces externos entre temas, de las revistas (85 revistas distintas) que han publicado artículos de dichos temas y en función de los investigadores (305 autores distintos) que han publicado artículos de los diferentes temas de investigación.

Se han utilizado técnicas de clasificación, "clustering", y de análisis de escalamiento multidimensional, "MDS". En ambas técnicas, las variables de estudio, han sido los 19 temas de investigación forestal obtenidos para el periodo 1994-2003 y los casos, los 19 temas de investigación, las 85 revistas o los 305 autores. El clustering realizado ha transformado las matrices asimétricas de entrada (revistas/temas y autores/temas) en matrices simétricas de similaridades (Pearson) y matrices de distancias (1 – Pearson) y ha agrupado los temas de investigación (método del vecino más lejano, "Complete Linkage") en un dendrograma horizontal. El análisis de escalamiento multidimensional ha permitido obtener, a partir de las matrices de distancias (1-Pearson), los archivos de coordenadas y los gráficos MDS. Con los archivos de coordenadas se han realizado en EXCEL gráficos de burbujas.

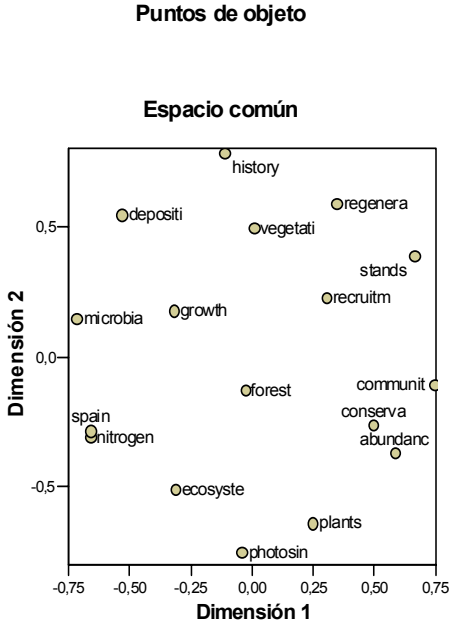
RESULTADOS

I. Análisis de los temas según la co-ocurrencia de descriptores

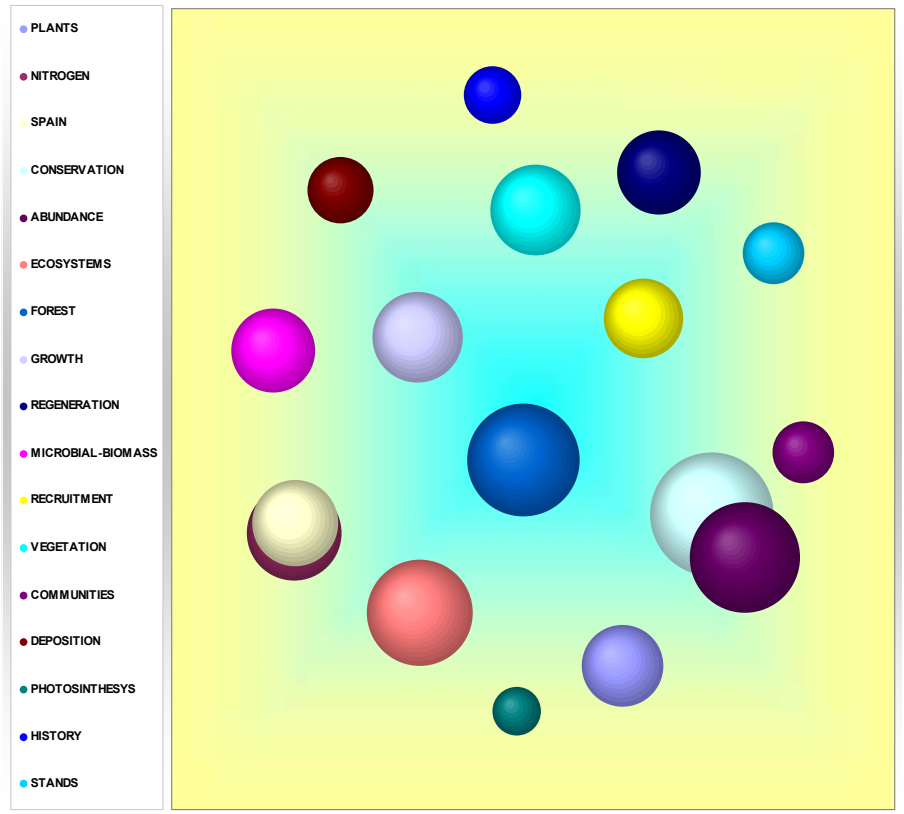
El agrupamiento de los temas de investigación obtenido según los enlaces externos entre ellos, ha sido el siguiente (SPSS 12.0):



El gráfico MDS (SPSS 12.0) obtenido ha sido el siguiente:

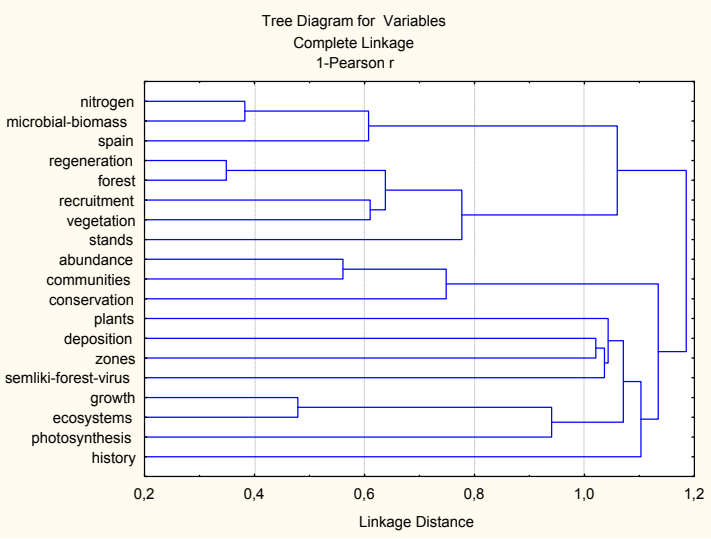


Llevadas las coordenadas del gráfico MDS a EXCEL, se ha obtenido el siguiente gráfico de burbujas (el tamaño del área se ha hecho proporcional al nº de artículos totales de cada tema)

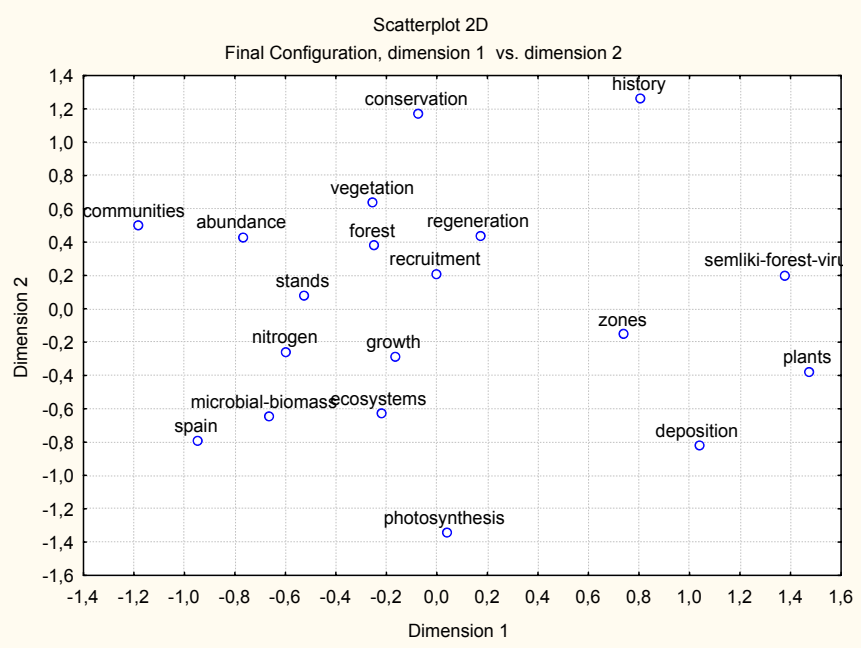


II. Análisis de los temas según las revistas

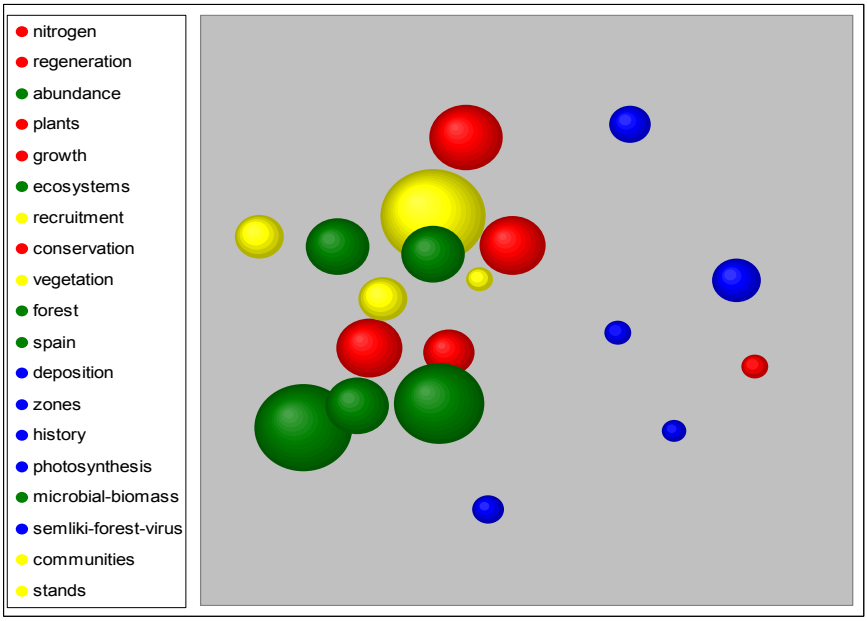
El agrupamiento de los temas de investigación obtenido según las revistas de publicación de los artículos de los diferentes temas, ha sido el siguiente (STATISTICA 6.0):



El gráfico MDS (STATISTICA 6.0) obtenido ha sido:



A partir de las coordenadas MDS se ha realizado el siguiente gráfico de burbujas en EXCEL (el tamaño de la burbuja se ha hecho proporcional al nº de artículos del tema publicados en la élite de revistas. Los colores representan: rojo (temas motor de investigación), verde (temas generales o estructurales), amarillo (temas nacientes o en proceso de desaparición) y azul (temas de disciplinas adyacentes):

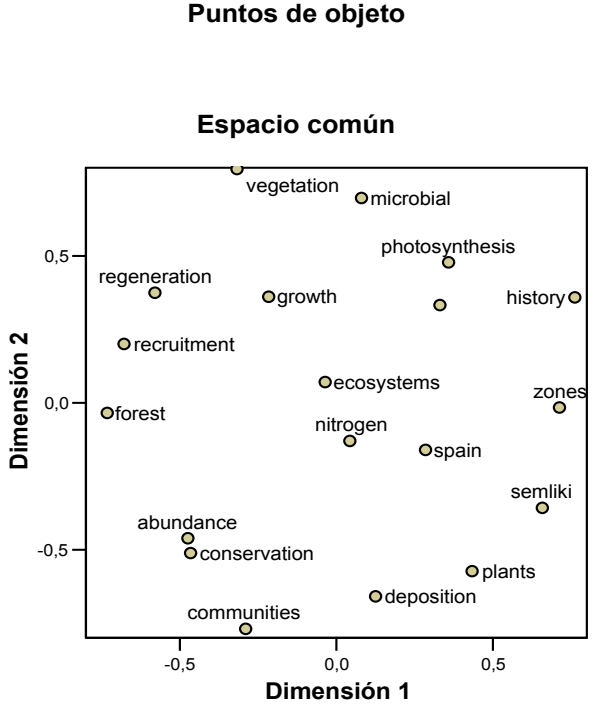


III. Análisis de los temas según los autores

El agrupamiento de los temas de investigación obtenido según la élite de autores, ha sido el siguiente (SPSS 12.0):

```
* * * * * * H I E R A R C H I C A L  C L U S T E R  A N A L Y S I S *
* * * * *

Dendrogram using Complete Linkage

                    Rescaled Distance Cluster Combine

  C A S E      0         5        10        15        20        25
 Label     Num  +---------+---------+---------+---------+---------+

abundanc     3
conserva     8
communit    19
regenera     2
forest      10
recruitm     7
nitrogen     1
ecosyste     6
spain       11
depositi    12
microbia    16
zones       13
semliki     17
photosyn    15
stands      18
plants       4
growth       5
history     14
vegetati     9
```

El gráfico MDS (SPSS 12.0) obtenido ha sido:



DISCUSIÓN

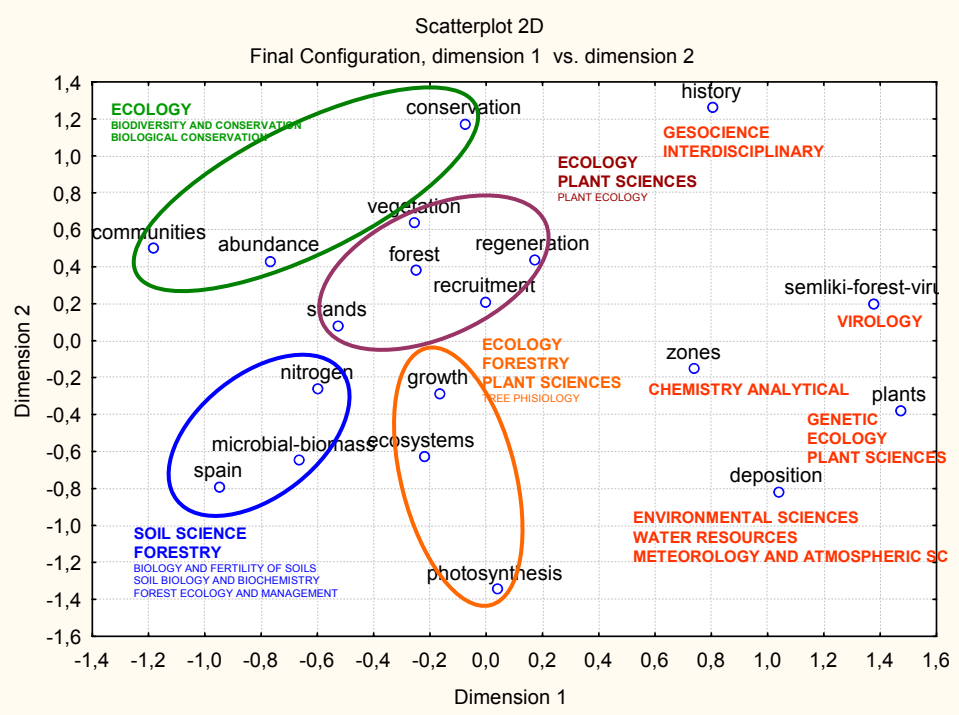
El análisis estadístico realizado sobre los temas de investigación ha puesto de manifiesto la existencia de los siguientes grupos significativos de temas:

Grupos	Análisis Temas/descriptores	Análisis Temas/revistas	Análisis Temas/autores
abundance-conservation	si (+++)	si(+)	si(++++)
abundance-communities	si(+)	si(+++)	si(+)
photosynthesis-stands	no	no	si(+++)
photosynthesis-plants	si(++)	no	si(+)
ecosystems-growth	si(+)	si(+++)	no
ecosystem-spain	no	no	si (+)
nitrogen-ecosystems	no	no	si(++)
nitrogen-spain	si(++++)	si(+++)	si (+)
nitrogen-microbial	si (+)	si(++++)	no
forest-regeneration	si (+)	si(++++)	si(+)
recruitment-vegetation	si(+)	si(+++)	no
abundance-conservation –commun	si(+)	si(++)	si(+)
photosynthesis-plants-stands	no	no	si(+)
growth-ecosystems-photosynthesis	no	si(+)	no
nitrogen-spain-microbial	si(+)	si(+++)	no
nitrogen-ecosystems-spain	no	no	si(+)
forest-regeneration-recruitment	si(+)	si(++)	si(+)
forest-regeneration-recruitment-vegetation	si(+)	si(++)	no
forest-regeneration-recruitment-vegetation-stands	no	si(++)	no

(++++ = muy alta vinculación, +++ = alta vinculación, ++ = vinculación, + = vinculación baja, no = nula vinculación)

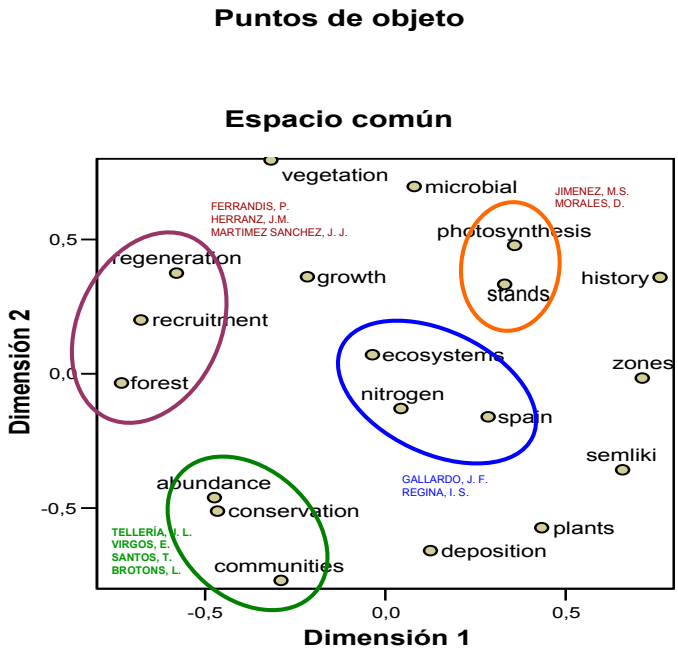
Como se puede observar en la tabla, los agrupamientos y su nivel de vinculación obtenidos para el conjunto de temas de investigación, varían según se hayan empleado como "casos" en los respectivos análisis, la co-ocurrencia de descriptores, las revistas o los autores. No se observa una coincidencia general en los grupos a nivel de existencia de vinculación, mucho menos sobre el grado de vinculación, tan sólo en los grupos ***nitrogen-spain, abundance-conservation-communities y forest-regeneration-recruitment*** (en la tabla en color amarillo), coinciden los tres análisis en la detección de vinculación significativa.

Si se tiene en cuenta que las técnicas de aglomeración agrupan las variables según la relación existente entre ellas, y que éstas van a venir dadas por los descriptores, las revistas o los autores comunes, se deduce que el empleo de una u otra relación será a elección del investigador según los objetivos del análisis a efectuar. Un agrupamiento de temas por revistas de publicación**, proximidad de publicación,** sería el indicado para conocer, por ejemplo, la clasificación de los temas de investigación según las categorías o disciplinas científicas de las revistas indexadas por ISI. Obsérvese el siguiente gráfico MDS:



Para cada grupo se indica la/s categoría/s ISI y las revistas comunes a los temas del grupo. Sombreados en azul aparecen los temas motor de investigación.

Así también, por ejemplo, el agrupamiento de autores**, proximidad de autoría o de autor,** sería de interés en los estudios de autoría y de grupos de investigación. Obsérvese el siguiente gráfico MDS:



Para cada grupo se indican los autores comunes a los temas de investigación.

O bién, por último, el agrupamiento de los temas de investigación según las relaciones semánticas existentes, **proximidad semántica**, sería de interés en estudios globales de

relaciones entre temas. Obsérvense los siguientes gráficos, EXCEL de burbujas y sociograma obtenido con el método de redes sociales:

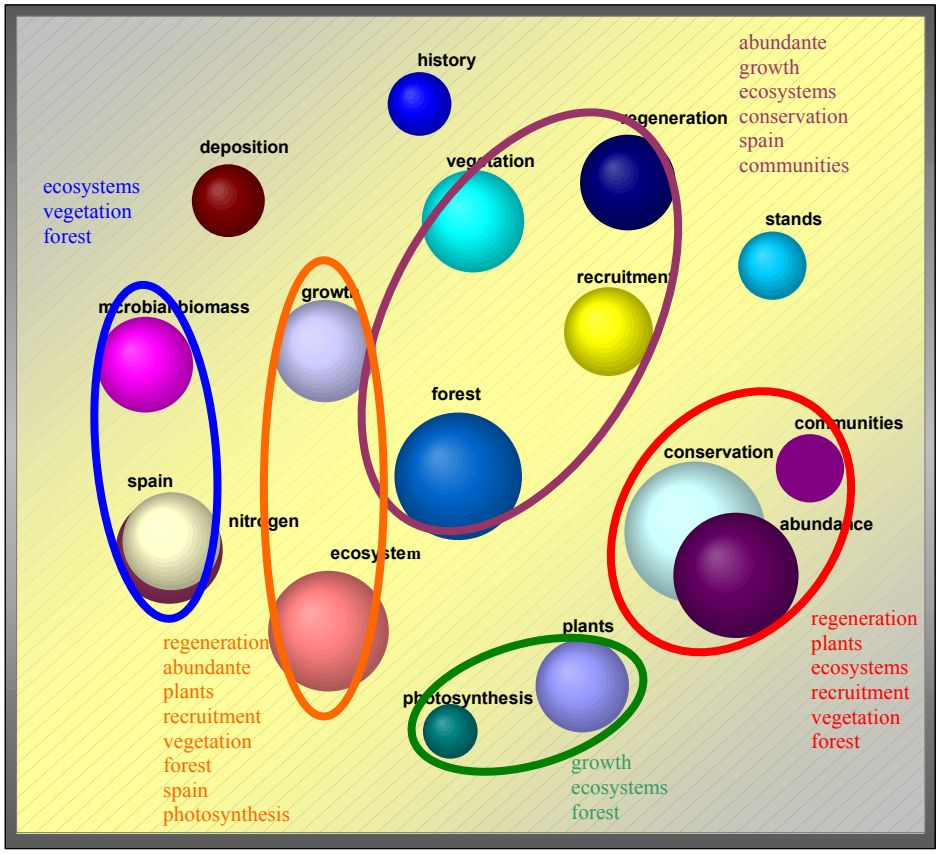
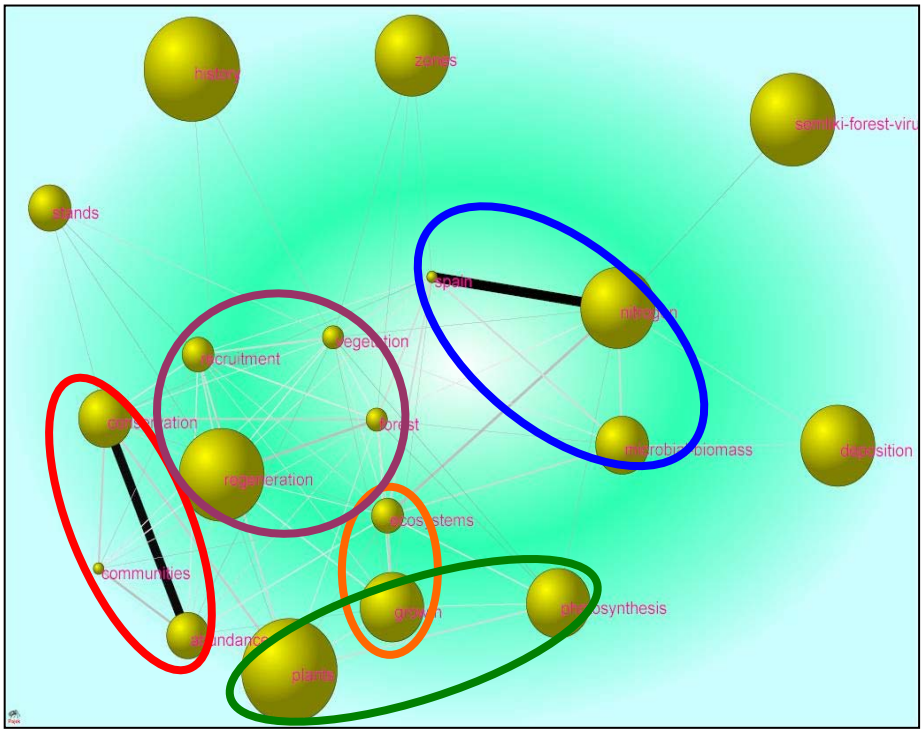


Gráfico de burbujas (EXCEL), obtenido a partir de las coordenadas MDS (SPSS12.0). Junto a cada grupo figuran los temas comunes con los que enlazan los temas de cada grupo.



Sociograma (PAJEK) (El tamaño de los nodos se ha hecho proporcional al grado de cohesión interna de los temas.

CONCLUSIONES

Las relaciones existentes entre un conjunto de variables, en nuestro caso temas de investigación, son diferentes según el conjunto de casos (enlaces externos con otros temas, revistas o autores) que se utilice para su agrupamiento. Según sea el objeto de estudio, al investigador le interesará conseguir agrupamientos en función de un conjunto u otro de casos. No obstante, cuando el objeto de estudio sean las propias relaciones existentes entre las variables, convendrá realizar todos los agrupamientos y observar los resultados obtenidos con todos los conjuntos de casos.

En el presente trabajo, tan sólo en los grupos ***nitrogen-spain, abundance-conservation-communities y forest-regeneration-recruitment***), se ha detectado coincidencia de vinculación significativa con los tres agrupamientos realizados. La existencia de vinculación significativa global, descriptores, revistas y autores, sería indicadora de pautas de investigación bien definidas, *"los mismos autores publican en las mismas revistas artículos sobre temas de investigación de semánticas poco diferenciadas"*. Así pues, la *cercanía total* entre temas de investigación vendria dada por la existencia de las tres proximidades, semántica, de publicación y de autoría. Un completo y exhaustivo análisis de las relaciones entre temas debería tener en cuenta, por tanto, los tres criterios de agrupación: semántico, de publicación y de autoría.

BIBLIOGRAFÍA

1. http://isiwebofknowledge.com/
2. ISI; 2002. Web of Science V5.0. Consultada 2005 enero 31. Disponible en: http://www.biblioteca.uclm.es/ayudas/webscience.htm
3. PINO-DÍAZ, J., RUIZ-BAÑOS, R. y BAILÓN MORENO, R.; 2005. *Producción científica forestal española, ISI 1994/2003*. (Artículo que se presenta al IV Congreso Forestal Español Septiembre-2005, Zaragoza)
4. CALLON, M., COURTIAL, J. P. & LAVILLE, F.; 1991. *Co-word analysis as a tool for describing the network of interactions between basic and technological research – the case of polymer chemistry.* Scientometrics 22 (1): 155-205.
5 JURADO-ALAMEDA, E., BAILÓN MORENO, R. y RUIZ-BAÑOS, R.; 2002. *Evaluación a través del análisis de las palabras asociadas. Aplicación a la evaluación de la investigación científica y técnica.* Ingeniería Química, marzo (141-149).
6. JURADO-ALAMEDA, E., BAILÓN MORENO, R. y RUIZ-BAÑOS, R.; 2002. *Evaluación a través del análisis de las palabras asociadas. Aplicación a la evaluación de la investigación científica y técnica.* Ingeniería Química, abril 177-184).
7. BAILÓN MORENO, R. *Ingeniería del conocimiento y vigilancia tecnológica aplicada a la investigación en el campo de los tensioactivos. Desarrollo de un modelo ciencimétrico unificado.* Consultada 2005 enero 31. Disponible en: http://www.mcu.es/cgi-bin/TESEO/BRSCGI2CMD=VERDOC&BASE=TSEO&DOCN=000093780
8. MOYA-ANEGÓN, F., VARGAS-QUESADA, B., HERRERO-SOLANA, V., CHINCHILLA-RODRIGUEZ, Z.; CORERA-ALVAREZ, E. y MUÑOZ-FERNÁNDEZ, F. J.; 2004. *A new technique for building maps of large scientific*

domains based on the cocitation of classes and categories. Scientometrics. Vol. 61, nº 1, 129-145.
9. http://vlado.fmf.uni-lj.si/pub/networks/pajek/
10. HANNEMAN, R. A.; *Introducción a los métodos de análisis de redes* sociales. http://www.redes-sociales.net/
11. MOLINA J.L., MUÑOZ, J., PHILIPPE, L.; *Red y realidad: aproximación al análisis de las redes científicas.* http://seneca.uab.es/antropologia/jlm/public_archivos/oviedo.pdf
12. MOLINA, J.L., MUÑOZ, J., DOMENECH, M.; *Redes de publicaciones científicas: un análisis de la estructura de coautorías.* http://seneca.uab.es/antropologia/jlm/public_archivos/copub.pdf
13. MOLINA, J.L.; 2001. *El análisis de redes sociales, una introducción*. Ed. Bellaterra, Barcelona. 124 p. ISBN: 84-7290-160-2
14. WHITE, H. D.; 2001. *Authord-centered bibliometrics through CAMEOs: Characterizations automatically made and edited online.* Scientometrics VOL. 51, No. 3, 607-637.
15. DE MOYA ANEGÓN, F., HERRERO-SOLANA, V. *Investigaciones en curso sobre interfaces gráficos en dos y tres dimensiones para el acceso a la información electrónica.* <http:://www.ucm.es/info/multidoc/multidoc/revista/num8/moya.html> [Consulta: 05/07/2005]
16. MARCOS, M.C. *Elementos visuales en sistemas de búsqueda y recuperación de información* [on line]. "Hipertext.net", núm. 3, 2005. <http://www.hipertext.net> [Consulta: 05/07/2005]. ISSN 1695-5498
17. HERRERO-SOLANA, V .*Técnicas avanzadas de recuperación de la información: Interfaces gráficos para la visualización de catálogos de bibliotecas.* http://www.ucol.mx/**interfaces**/**interfaces**2001/**interfaces**2000/ponencias/ponencias/Pm04.pdf [Consulta: 05/07/05]
18. DE LA ROSA**,** F.; POZO, S.; GASCA, R.M. (2005). *"Análisis y visualización de las comunidades científicas con información extraída de la web"*, Revista IEEE América Latina . Volume: 3, Issue: 1, March 2005. ISSN 1548-0992
19. PINO-DIAZ, J. (2005*). "Análisis de redes aplicado a la producción científica forestal española, ISI-1994/2003".* <http://perso.wanadoo.es/josepinodiaz/archivos_index/ART%CDCULO_DOCUMENTO%20HIPERTEXTO.htm> [Consulta: 05/07/05]