VEGANA, un paquete de programas para la gestión y análisis de datos ecológicos

Miquel De Cáceres[1](mcaceres@bio.ub.es), Xavier Font[1](xavier@bio.ub.es), Ricard García[1] y Francesc Oliva[2](francesc@bio.ub.es).

[1]Departamento de Biología Vegetal (Universidad de Barcelona). Avda. Diagonal 645, 08028 Barcelona, España
[2]Departamento de Estadística (Universidad de Barcelona). Avda. Diagonal 645, 08028 Barcelona, España

Resumen: VEGANA (Vegetation edition and Analysis) es un paquete integrado de programas destinado a la gestión y análisis de datos ecológicos en general y muy especialmente de vegetación. Todo el *software* de VEGANA está desarrollado en lenguaje Java, por lo que es posible su ejecución en todas las plataformas que soportan la máquina virtual de Java (JVM, http://java.sun.com/j2se/). Contiene 4 programas principales:

-Ginkgo: está orientado a la representación y clasificación de individuos a partir de datos multivariantes. Entre otras técnicas, permite realizar grupos jerárquicos y partitivos (*K-means* y *fuzzy C-means*); análisis discriminante lineal, cuadrático y basado en distancias; reducción de la dimensionalidad, análisis de componentes principales (PCA), análisis de coordenadas principales (MDS), *multidimensional scaling* no métrico (NMDS) y análisis factorial de correspondencias (CA).

-Quercus: es un editor de tablas de datos del tipo inventario. Permite la introducción y gestión de inventarios de especies y su posterior exportación a Ginkgo. Es posible también editar los ficheros tesauros asociados (taxones, comunidades y bibliografía). Los datos pueden ser importados del Banco de datos de biodiversidad de Cataluña (http://biodiver.bio.ub.es/biocat/homepage.html) mediante el formato XML.

-Fagus: es un editor y gestor de citas florísticas. Se pueden almacenar datos inéditos, bibliográficos y de colecciones.

-Yucca: permite la representación cartográfica de los datos, principalmente distribuciones de taxones o grupos de áreas.

El *software* presentado tiene una distribución libre y gratuita y la descarga del programa como las actualizaciones se realizan automáticamente gracias a la tecnología Java Web Start. La página principal de VEGANA se encuentra en http://biodiver.bio.ub.es/vegana/

Palabras clave: Análisis Discriminante, Análisis Multivariante, Editor de Tablas de Inventarios, Fitosociología, Software.

Introducción

VEGANA (VEGetation edition and ANAlysis) es un paquete de programas destinado a la gestión y análisis de datos ecológicos en general y, muy especialmente, de vegetación. En su conjunto, pretende ser una herramienta integrada de recopilación, manipulación y análisis de información ecológica. Contiene 4 programas, que pueden ser ejecutados independientemente (fig. 1):

a) *Ginkgo*, orientado a la representación y clasificación de individuos a partir de datos multivariantes.
b) *Quercus*, un editor de tablas de datos del tipo inventario. Permite la introducción y gestión de inventarios de especies y su posterior exportación a *Ginkgo*.
c) *Fagus*, un editor y gestor de citas florísticas. Permite almacenar principalmente datos inéditos, bibliográficos o procedentes de colecciones de pliegos de herbario.
d) *Yucca*, permite la representación cartográfica de los datos, principalmente distribuciones de taxones, comunidades o grupos de áreas.

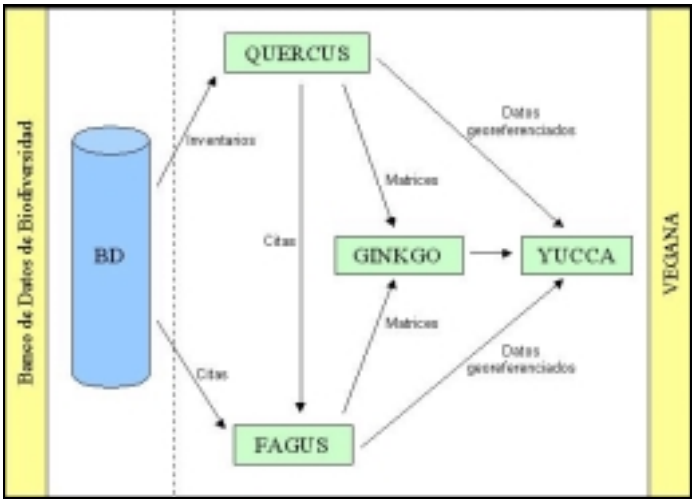

Figura 1: Esquema de los módulos de VEGANA y sus relaciones.

Características técnicas

Todo el *software* de VEGANA esta desarrollado en lenguaje Java (ver. 1.4.x). Gracias a ello, es posible su utilización en un gran numero de sistemas operativos (Windows, Linux, Mac, etc...). En la práctica es posible su utilización en cualquier plataforma que soporte la máquina virtual de Java (JVM, http://java.sun.com/j2se).

La página principal del paquete se encuentra en: http://biodiver.bio.ub.es/vegana, donde es posible la descarga de los programas y ficheros de tesauros asociados. La distribución de VEGANA es gratuita y las actualizaciones del programa se realizan automáticamente gracias a la tecnología Java Web Start. Los requerimientos mínimos de hardware que estimamos son los de un procesador Pentium III con una memoria igual o superior a 256 Mb.

Tanto los ficheros de datos como los ficheros de configuración de los programas se guardan en formato XML (http://www.xml.org), por lo que pueden ser visualizados externamente mediante programas que soportan este estándar.

Tesauros

Dado que QUERCUS y FAGUS son editores de datos asociados a taxones, es conveniente asociar dichos datos a un fichero tesauro. Esto presenta tres ventajas: En primer lugar, los nombres de los taxones pueden ser comprobados evitando errores tipográficos en la introducción de datos. En segundo lugar, permite dar soporte a los nombres sinónimos, facilitándose así el posterior análisis de datos al interpretarse todos los nombres sinónimos como un único taxón. En tercer lugar, los tesauros pueden almacenar información biológica de las especies, cosa que hace posible la elaboración de espectros biológicos (formas biológicas, distribución, etc.). Así pues, para usar estos dos editores es necesario disponer de un tesauro de taxones. Desde la página web de VEGANA se proporciona actualmente un tesauro para la flora catalana y otro para la flora europea. Es posible el uso y gestión de tesauros sintaxonómicos y bibliográficos, aunque ello no es imprescindible para la correcta utilización del programario. QUERCUS y FAGUS proporcionan las herramientas necesarias para generar y actualizar todos los ficheros de tesauro (fig. 2).

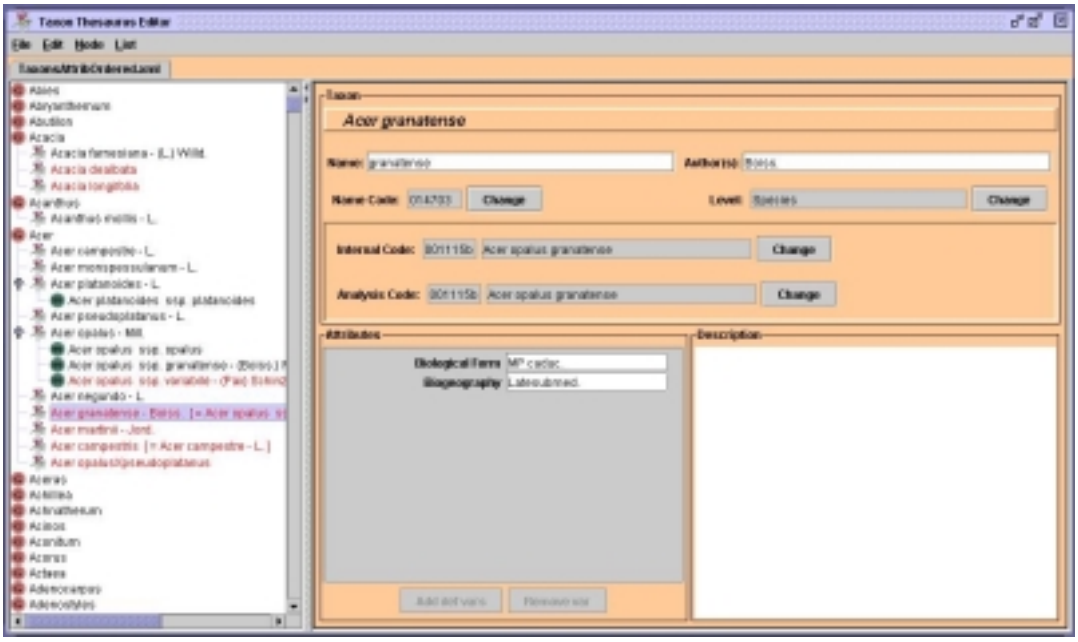

Figura 2: Aspecto del editor de tesauros de taxones, disponible en QUERCUS y FAGUS.

Ficheros de configuración

Todos los programas funcionan a través de ficheros de proyectos. Un proyecto contiene la lista de recursos (tesauros) y ficheros de datos utilizados en el programa, los campos definidos por el usuario, así como la configuración general del programa (opciones visuales, idioma, ...). El formato del fichero de proyecto es, como en el caso de los ficheros de datos, también XML.

GINKGO

GINKGO acerca varias herramientas de análisis multivariante a usuarios no expertos en estadística, a través de una interfaz gráfica sencilla. Dicha interfaz de usuario presenta 3 ventanas flotantes: el editor de datos, el gestor de análisis y el gestor de gráficos (fig. 3). Éstas proporcionan un marco integrado de trabajo que permite la exploración paso a paso de datos multivariantes. En primer lugar, permite elegir entre distintos coeficientes de similaridad y disimilaridad adecuados a los datos de los que se dispone. A continuación, la estructura de los datos puede ir siendo dilucidada empleando las distintas técnicas de reducción de la dimensionalidad y clasificación incluidas en el programa. Finalmente, los resultados obtenidos con el empleo de distintos espacios y/o

técnicas de análisis pueden ser luego comparados en el mismo programa. Cabe resaltar, que es sencillo traspasar de nuevo al editor de datos las matrices generadas en los análisis, con lo que se pueden realizar procesos de análisis moderadamente complejos. El programa permite guardar tanto matrices de datos como resultados de análisis en un solo fichero de proyecto, con lo que un proceso largo de análisis de datos puede realizarse en varias sesiones.

Dado que en los proyectos de GINKGO se almacenan las matrices de datos y resultados, esto generaría ficheros XML de proyecto muy grandes. Por esta razón, el fichero del proyecto en este programa se puede guardar opcionalmente de forma comprimida.

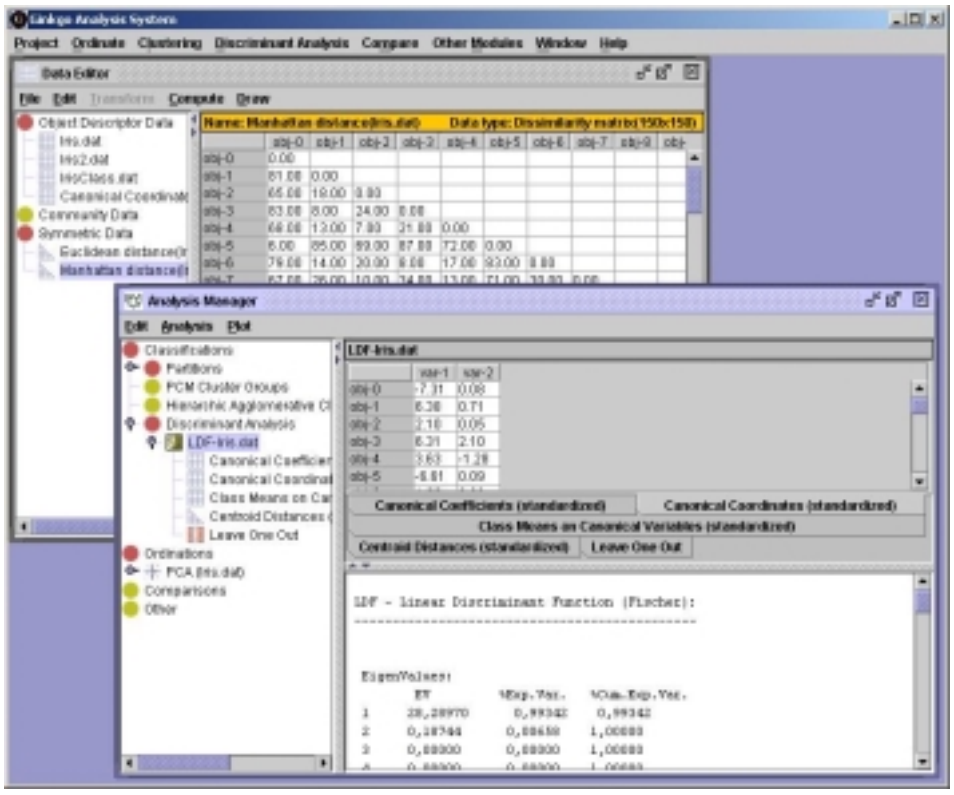

Figura 3: Aspecto general del programa GINKGO.

Edición de datos multivariantes

El editor de datos de GINGKO permite una edición directa de matrices de datos. Se diferencian dos tipos de matrices, rectangulares (objeto-descriptor) y simétricas (objeto-objeto o descriptor-descriptor). Ambos tipos de matrices de datos se pueden crear dentro

del programa aunque también se permite la importación de datos en formato de texto ASCII delimitado por tabuladores. La exportación de tablas de resultados hacia otros programas se puede hacer con el mismo formato o bien a través del portapapeles. Además de los estadísticos descriptivos univariantes, el programa ofrece otras operaciones comunes, como son la estandarización de variables, la transposición de matrices o el cálculo de matrices de covariancia/correlación.

Técnicas de ordenación

GINKGO permite el uso de varias técnicas clásicas de reducción de la dimensionalidad:

- Análisis de Componentes Principales (*PCA*)
- Análisis de Coordenadas Principales (o *Metric scaling*, Gower, 1966)
- *Multidimensional scaling* No Métrico (*NMDS*, Kruskal 1964a, 1964b).
- Análisis Factorial de Correspondencias (*CA*, Hill 1973).

Técnicas de clustering

En cuanto a técnicas de *clustering* se refiere, GINKGO permite aplicar 3 modelos distintos de clasificación:

a) *Clustering* Jerárquico Aglomerativo. Se permite la selección entre los algoritmos *Single Linkage*, *Complete Linkage*, *UPGMA*, *WPGMA*, *UPGMC*, *WPGMC*, método de Ward, *Flexible clustering*. Cabe resaltar que el programa permite "cortar" un dendrograma por el nivel de similaridad deseado para producir particiones que pueden ser comparadas con el uso de otras técnicas de clasificación.
b) Algoritmos partitivos: *K-means* (MacQueen, 1967) y *Fuzzy C-means* (*FCM*, Bezdek, 1981), generalización de K-means al enfoque basado en la lógica borrosa.
c) *Clustering* no partitivo: *Possibilistic C-means* (*PCM*, Krishnapuran y Keller, 1993, 1996) surge de la relajación del concepto de partición. Con este modelo de clasificación cada *cluster* se determina independientemente comparando la distancia del objeto al centroide con una distancia de referencia.

Es un problema habitual en usuarios de aplicaciones de estadística multivariante el estar limitado al uso de la Distancia Euclídea en los paquetes estadísticos convencionales. Esto es así debido a la carencia de otras medidas de distancia en dichos paquetes y al

empleo implícito o explícito de la Distancia Euclídea dentro de las mismas técnicas multivariantes que los programas comunes ofrecen. Los usuarios avezados a este problema acostumbran a transformar previamente los datos antes de empezar el análisis o emplear técnicas de *scaling*. No obstante, muchos usuarios inexpertos se conforman con lo que el paquete estadístico les ofrece. GINKGO intenta suplir estas carencias de varios modos. En primer lugar, incorpora medidas de similaridad y distancia poco comunes en los paquetes estadísticos estándar y que son frecuentemente utilizadas en ámbitos de aplicación, como en ecología o taxonomía numérica. En segundo lugar, GINKGO ofrece la posibilidad de transformar matrices de similaridad en disimilaridad y viceversa. Finalmente, proporciona técnicas de ordenación y clasificación aplicables a matrices de disimilaridades, conservando las propiedades de la métrica usada íntegramente en los análisis posteriores. Concretamente, es posible ejecutar los algoritmos de *clustering* como *K-means* y *FCM* a partir de una matriz de disimilaridades cualquiera (Oliva et al., 2001), opción raramente disponible en programas comerciales.

Otra característica que GINKGO ofrece y es poco común en otros programas, es permitir la utilización de forma inmediata de los resultados de *clustering* para distinguir los grupos en las representaciones gráficas de dimensionalidad reducida (fig. 4). De este modo, la interpretación de ambas técnicas de análisis se pueden complementar y contrastar con facilidad.

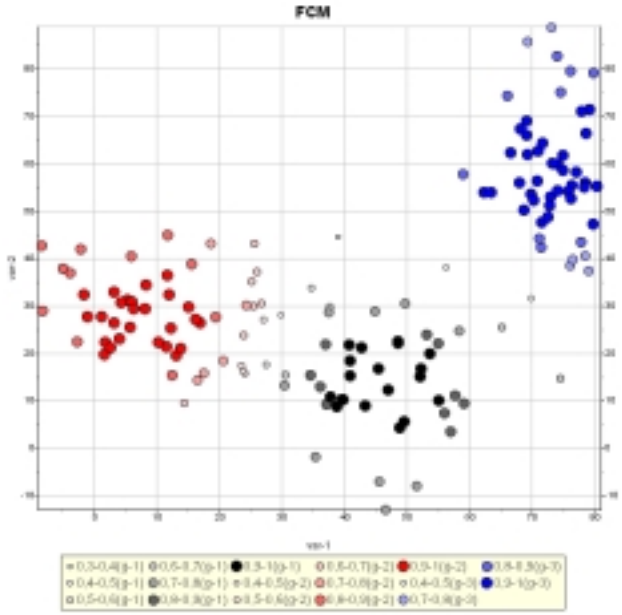

Figura 4: Diagrama de dispersión con la visualización de conjuntos borrosos.

Análisis Discriminante

Las técnicas de análisis discriminante actualmente disponibles en GINKGO son:

a) Análisis Discriminante Lineal (Canónico). Entre las opciones disponibles se encuentran elegir el número de ejes canónicos y la normalización de los vectores propios.
b) Análisis Discriminante Cuadrático.
c) Análisis Discriminante Basado en Distancias (Cuadras et al., 1997). Análogamente al caso de *k-means*, podemos emplear cualquier medida de disimilaridad para un análisis discriminante. La matriz de entrada para este tipo de análisis es una matriz simétrica de disimilaridades entre objetos.

Comparación de Particiones

Es posible en GINKGO comparar las particiones surgidas de distintos análisis entre ellas o con clasificaciones externas. Dicha comparación se realiza mediante el índice de Rand corregido (Rand, 1971; Hubert y Arabie, 1980) y la confección de matrices de confusión.

QUERCUS

QUERCUS se ha concebido como una herramienta de edición, manipulación y almacenaje de inventarios de vegetación. Originalmente, se ha diseñado para desarrollar trabajos fitosociológicos, pero también puede ser útil para otro tipo de estudios ecológicos que se basen en la manipulación de matrices de datos sobre especies. Como valores de abundancia de taxones, admite tanto los coeficientes de la escala de Braun-Blanquet como porcentajes de cobertura o simplemente presencia/ausencia de taxones.

QUERCUS surge de la evolución de otro programa para gestionar inventarios de vegetación desarrollado en nuestro grupo, llamado XTRINAU (Font, 1990; Font y Ninot, 1995), operativo únicamente bajo el sistema DOS. Aunque QUERCUS es básicamente un editor, permite funciones parecidas a las que realizamos en una base de datos. Es posible definir nuevos campos específicos para los inventarios y hacer búsquedas sobre ellos.

Edición de Tablas de Inventarios Primarias

Las tablas primarias contienen inventarios, aún sin elaborar desde el punto de vista taxonómico y sintaxonómico. Para la definición de campos a considerar en los inventarios se han seguido las directrices propuestas por Mucina et. al (2000). El conjunto de tablas primarias forman el banco de inventarios del autor y pueden presentar tanto inventarios inéditos como publicados. El Editor de Tablas Primarias de QUERCUS permite informatizar y almacenar inventarios de una manera sencilla y guiada. Presenta tres formas de edición de tablas de inventarios: 1) Edición de toda la tabla de entradas de taxones a la vez, 2) edición de cada inventario por separado y 3) edición de la información asociada al inventario en forma de tabla.

Es posible importar tablas de inventarios desde el Banco de Datos de Biodiversidad de Cataluña (Font et. al, 2001) directamente a QUERCUS. La página principal del Banco de Datos se encuentra en http://biodiver.bio.ub.es/biocat/homepage.html.

Manipulación de Tablas de Trabajo (o Secundarias)

A partir de los inventarios almacenados en las tablas primarias se elaboran las tablas de trabajo o secundarias. Dichas tablas de trabajo son las que el fitosociólogo clásico manipula para establecer unidades de vegetación. En el caso de QUERCUS, el Editor de Tablas de Trabajo (fig. 5) permite la preparación de la tabla para ser exportada para la publicación o para ser analizada por metodos numéricos. Algunas de operaciones que se realizan de manera automática con este editor son el filtrado o suma de entradas de taxones según condiciones específicas, como el estrato en que aparecen las plantas o la certeza de la determinación.

La homogeneización nomenclatural es otro aspecto que toma importancia en la preparación de una tabla para ser analizada. Se distinguen al menos 3 situaciones básicas: 1) Eliminar los sinónimos que puedan aparecer en los distintos inventarios y añadir los valores de abundancia al taxón válido. 2) Como no todos los autores afinan de igual modo sus determinaciones infraespecíficas debemos considerar estos taxones

en el sentido más amplio. 3) Podemos encontrar taxones polimórficos de difícil determinación. La primera situación se resuelve con un tesauro que permita reconocer los distintos nombres sinónimos de un taxón. Los otros dos casos requieren el uso de un segundo nivel de sinonimia, orientada al análisis, que permita tratar conjuntamente dos taxones ambos correctos desde un punto de vista nomenclatural i taxonómico. El editor de tablas de trabajo permite resolver estos 3 casos, empleando para ello el tesauro de taxones.

También es posible realizar análisis simples de la estructura de la comunidad, mediante la confección de espectros biológicos basados en la corología de los taxones, su tipo biológico o cualquier otro parámetro biológico que anteriormente se haya informatizado en el tesauro de taxones.

Finalmente, el editor permite exportar una tabla de inventarios a una matriz de datos numéricos, susceptible de ser analizada con métodos multivariantes. En el caso de valores de abundancia medidos en la escala de Braun-Blanquet, es necesario transformar los valores a cantidades numéricas siguiendo una tabla de equivalencias. La matriz resultante puede ser usada en GINKGO o almacenada en formato ASCII para su uso en otros programas.

Tablas Sintéticas

Una tabla sintética es esencialmente como una tabla de inventarios en la que cada columna corresponde a las frecuencias de aparición de los taxones presentes en un grupo seleccionado de inventarios. La creación de tablas sintéticas se hace, pues, únicamente a partir de tablas de trabajo. Si se dispone de un tesauro sintaxonómico, se pueden crear tablas sintéticas siguiendo la asignación fitosociológica de los inventarios. Los valores de los inventarios son valores constantes expresados en porcentaje o clases de porcentajes. El Editor de Tablas Sintéticas permite manipular dichas tablas sintéticas. Muchas de las operaciones del editor de tablas de trabajo se hallan también disponibles aquí.

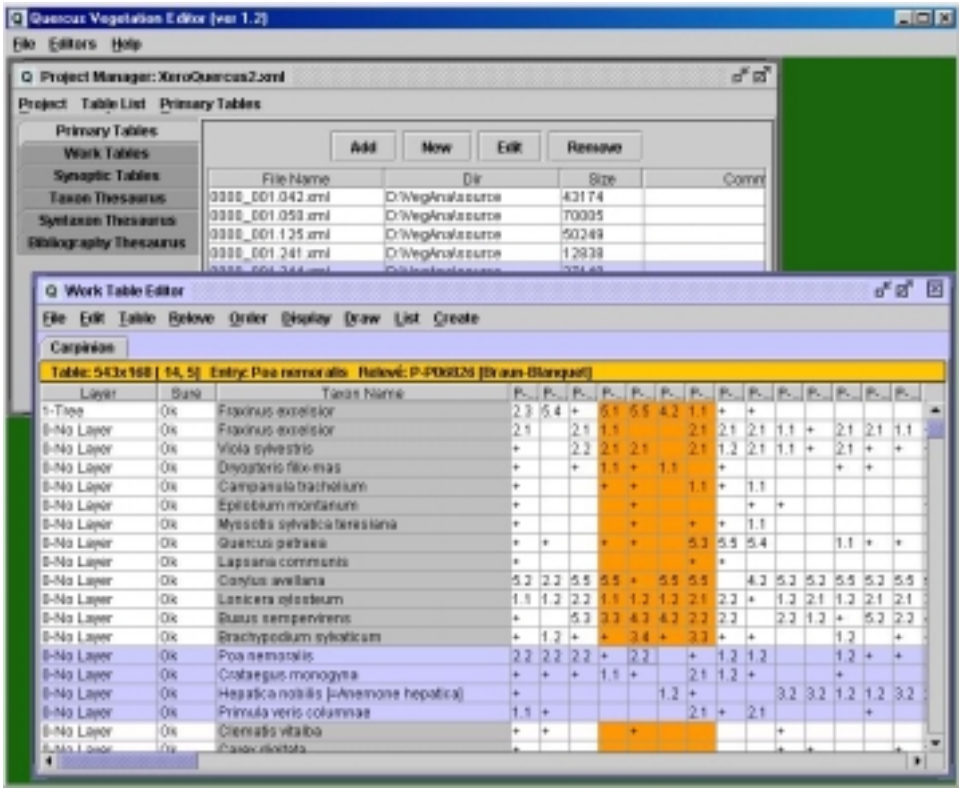

Figura 5: Aspecto general de la interfaz de QUERCUS con el editor de tablas de trabajo.

FAGUS

Esta aplicación permite informatizar y recopilar citas florísticas, para luego elaborarlas y generar floras, etiquetas, distribuciones, etc. Dichas citas pueden ser tanto pliegos de herbario como muestras de campo o citas bibliográficas.

Tablas de Citas

FAGUS permite la confección sencilla y guiada de listas (tablas) de citas florísticas (fig. 6). En cada cita se almacenan datos básicos como el nombre del taxón (relacionado en el tesauro de taxones), el nombre del autor y la fecha de la observación, la georeferenciación, etc. Además, es posible asociar a la cita taxonómica nuevos campos definidos por el usuario, permitiendo así que cada usuario pueda configurar el programa según sus propias necesidades. Como en QUERCUS, este módulo también puede importar citas procedentes del Banco de Datos de Biodiversidad de Cataluña (http://biodiver.bio.ub.es/biocat/homepage.html).

Además de la confección y almacenamiento de tablas de citas, FAGUS presenta también otras utilidades que permiten tratar dichas tablas. Empleando el ”Buscador de Citas” se pueden agrupar las citas de diferentes tablas según el valor de un campo determinado, escogido por el usuario. Ello permite crear tablas nuevas con, por ejemplo, todas las citas de un taxón concreto o todas las citas localizadas en un cierto cuadrado UTM.

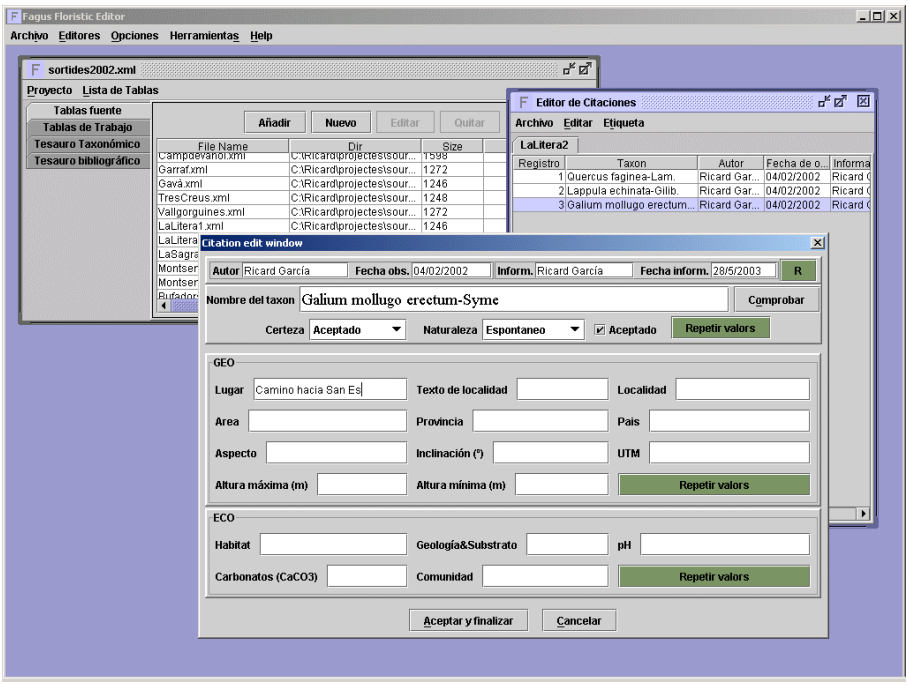

Figura 6: Aspecto general de la interfaz de FAGUS con el panel de informatización de citas.

Generación de Floras

Mediante el “Generador de Flora” se puede confeccionar una flora con todas las citas almacenadas en FAGUS. El usuario puede elegir qué tablas quiere incluir en confección de la flora, así como qué tipo o tipos de citas deben aparecer y qué datos se presentarán de cada una de ellas. Los taxones se pueden ordenar alfabética o taxonómicamente. La flora generada se guarda en un fichero RTF.

Etiquetas de herbario

FAGUS permite asociar al archivo proyecto la imagen de una etiqueta en la cual imprimir los datos informatizados y confeccionar las etiquetas de herbario. Es posible

también variar el tamaño de impresión de la etiqueta o el tamaño del texto que se escribirá sobre ésta.

YUCCA

Este programa es una herramienta de visualización cartográfica de información georeferenciada. Se puede generar mapas de distribuciones de taxones, sintaxones, abundancia de citas florísticas, etc (fig. 7). Para ello es necesario disponer de imágenes de mapas en las que dibujar los símbolos así como calibrar las imágenes con puntos de referencia. En la página de VEGANA se pueden obtener configuraciones de mapas para las áreas geográficas catalana y peninsular. Ambos editores, QUERCUS y FAGUS, permiten exportar sus datos para ser visualizados en YUCCA.

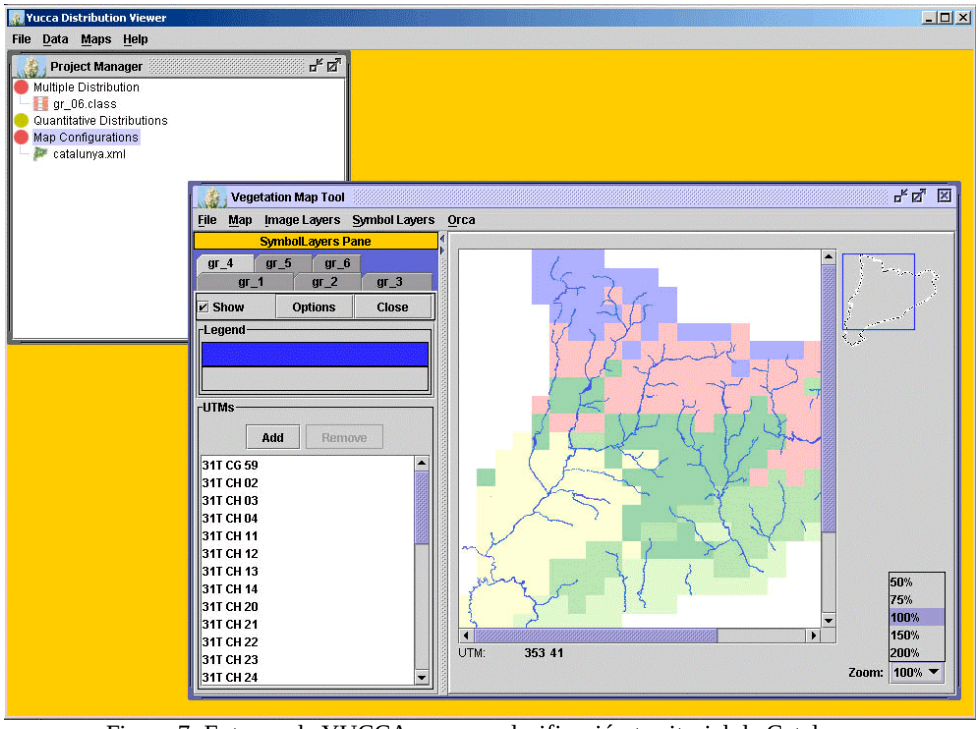

Figura 7: Entorno de YUCCA con una clasificación territorial de Catalunya.

Agradecimientos

El presente trabajo se ha realizado con el soporte del “Comissionat per a Universitats i Recerca” (1999SGR00059), del “Departament d’Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya” (2001 FI 00269) y gracias al Convenio de colaboración entre el Departamento de Medio Ambiente de la Generalitat de Catalunya y la Universidad de Barcelona para la elaboración de información sobre la biodiversidad y el patrimonio natural de Catalunya.

Referencias bibliográficas

Bezdek, J.C. (1981). Pattern recognition with fuzzy objective functions. Plenum Press. New York.

Cuadras, C.M., Fortiana, J. y Oliva, F. (1997): The proximity of an individual to a population with applications in discriminant analysis. Journal of Classification 14, 117-136.

Font, X. (1990): XTRINAU (ver. 1.0). Un programa para la gestión de los inventarios fitocenológicos. Monografias del Instituto pirenaico de Ecología 5: 531-539. Jaca.

Font, X. y Ninot, J.M. (1995): A regional project for drawing up inventories of flora and vegetation in Catalonia (Spain). Ann. Bot. (Roma) 53: 99-105.

Font, X., Cáceres, M. de, y Quadrada, R. (2001): La biodiversitat de Catalunya consultable via Internet, http://biodiver.bio.ub.es/biocat/homepage.html. L'Atzavara, butlletí de la Sec. de Ciències Naturals del Museu de Mataró 9: 57-58.

Gower, J.C. (1966): Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53, 325-338.

Hill, M. O. (1973): Reciprocal averaging: An eigenvector method of ordination. Journal of ecology 61, 237-249.

Hubert, L. y Arabie, P. (1985) Comparing partitions. Journal of Classification 2: 193-218.

Krishnapuram, R. y Keller, J.M. (1993): A possibilistic approach to clustering. IEEE transactions on fuzzy systems 1, 98-110.

Krishnapuram R. y Keller J.M. (1996): The possibilistic c-means algorithm: Insights and recommendations. IEEE transactions on fuzzy systems 4, 385-393.

Kruskal, J.B. (1964a): Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. 29(1), 1-27.

Kruskal, J.B. (1964b): Non-metric Multidimensional scaling: A numerical method. Psychometrika 29(2), 115-129.

MacQueen, J. (1967): Some methods for classification and analysis of multivariate observation. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, pp. 281-297.

Mucina, L., Joop, H.J.S. y Rodwell, J.S. (2000): Common data standards for recording relevés in field survey for vegetation classification. Journal of Vegetation Science 11: 769-772.

Oliva, F., De Cáceres, M., Font, X. y Cuadras, C.M. (2001): Contribuciones desde una perspectiva basada en distancias al fuzzy C-means clustering. XXV Congreso Nacional de Estadistica e Investigación Operativa. Úbeda, 2001.

Rand, W.M. (1971): Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association 66: 846-850.