CONOCIMIENTO ABIERTO EN SISTEMAS DE INFORMACIÓN GEOGRÁFICA. UNA ESTRATEGIA PARA LA GEOGRAFÍA FÍSICA

Francisco Alonso Sarría[*], Francisco Gomariz Castillo[**], Fulgencio Cánovas García[***]
Universidad de Murcia

RESUMEN
La aparición de un nuevo paradigma científico basado en un uso intensivo de al informática afecta a las ciencias ambientales y también a la Geografía, y en particular a la Geografía física.
En este nuevo paradigma, datos, programas y algoritmos pasan a tener un papel central. Diversos autores afirman la necesidad de utilizar programas abiertos y publicar los algoritmos utilizados junto con los resultados obtenidos.
Este trabajo propone un entorno computacional para el trabajo en Geografía física, basado en software abierto y con capacidad para la programación de algoritmos. Los programas incluidos son bien conocidos por la comunidad científica y ampliamente utilizados.
Palabras clave: Ciencia intensiva en datos, SIG, Software de código abierto, Geografía Física

Open Knowledge in Geographical Information Systems. A Strategy for Physical Geography

ABSTRACT
The emergence of a new scientific paradigm based on an intensive use of computers is affecting environmental sciences and Geography as well, specially in the case of Physical geography.
In this new paradigm, data, programs and algorithms acquire a central role. Several authors have claimed the need to use open software and to publish the algorithms used to obtain published results.
This paper proposes a computational environment for working in Physical Geography. It si based on open software and includes programming capabilities. The programs are well known by the scientific community and are widely used.
Keywords: Data-intensice science, GIS, Open source software, Physical Geography

1. INTRODUCCIÓN

La creciente importancia de la informática en la investigación científica ha llevado a algunos autores (Hey et al., 2009a) a postular la aparición de un nuevo paradigma en la ciencia.

Fecha de recepción: 9 de mayo de 2012
Fecha de aceptación: 9 de julio de 2012
[*] Departamento de Geografía. Instituto Universitario del Agua y del Medio Ambiente. Universidad de Murcia. E-mail: alonsarp@um.es
[**] Instituto Euromediterráneo del Agua. E-mail: fjgomariz@um.es
[***] Instituto Universitario del Agua y del Medio Ambiente. Universidad de Murcia. E-mail: fulgencio.canovas@um.es

El primer paradigma fue el de la ciencia experimental al que siguió el de la *ciencia teórica*. Cuando los modelos de esta se hicieron demasiado complejos para ser resueltos analíticamente, empezaron a resolverse mediante simulaciones. Se genera así el tercer paradigma, la *ciencia computacional*.

Estas simulaciones habrían acabado por generar una gran cantidad de nuevos datos que, junto con los que sigue produciendo la ciencia experimental, plantean nuevos retos en cuanto a su localización, organización y análisis. Las tecnologías y técnicas necesarias para organizar una investigación con un uso tan intensivo de datos son, según estos autores, tan diferentes a las utilizadas anteriormente, que creen justificado hablar de un cuarto paradigma que será el de la *ciencia intensiva en datos*[1] (Bell et al., 2009).

El uso intensivo de la informática, bien como computación o bien en el seno de este nuevo paradigma, ha supuesto un cambio fundamental en la manera de hacer ciencia (Mesirov, 2010; McCafferty, 2010; Ince et al., 2012).

Una de las disciplinas que, de forma reciente, se está viendo más afectada por estos cambios es la Ecología (Hunt et al., 2009; Reichman et al., 2011; Michener y Jones, 2012) y, en general, las ciencias ambientales, debido al enfoque global y multidisciplinar que utilizan.

Esta nueva situación se refleja en la aparición de nuevos términos. En Ecología aparece la *Ecología computacional*, basada en simulaciones y la *Ecoinformática*, orientada a la recolección y análisis de grandes cantidades de información. Conceptos similares (*Biología computacional y Bioinformática*) aparecen en el campo de la Biología (Hey et al., 2009b) y en otras ciencias vinculadas a la Geografía física, por ejemplo *Hidroinformática* (Kumar et al., 2005), implicando una Hidrología intensiva en datos, o *Pedometría/Cartografía digital de suelos* (Lagacherie et al., 2007; Boettinger et al., 2010) que puede entenderse más como Edafología computacional.

Si consideramos a la Geografía física como una ciencia ambiental con un enfoque pluridisciplinar (Tricart y Kilian, 1979; Stoddart, 1986; Goudie, 1994; Sala y Batalla, 1996; Slaymaker y Spencer, 1998; Gregory, 2000; Holden, 2008), resulta evidente que se encuentra plenamente afectada por este nuevo paradigma. En este trabajo se analizarán algunas de sus consecuencias en relación al uso de la informática, se justificará su relevancia en Geografía física y se propondrá un entorno de trabajo computacional, basado en software abierto, que pudiera resultar útil para la Geografía física.

2. GEOGRAFÍA Y GEOGRAFÍA FÍSICA

La Ecología es una ciencia sintética que se beneficia del acceso abierto a datos procedentes de las ciencias de la Tierra, la vida y las ciencias sociales (Reichman et al., 2011). Por otra parte, se convierte cada vez más en una ciencia intensiva en el uso de datos: teledetección, redes de sensores, redes de observatorios, etc. (Michener y Jones, 2012).

Los cambios que están ocurriendo en la Ecología crean desafíos con respecto a la adquisición, gestión y análisis de grandes volúmenes de datos que son recogidos por los científicos en todo el mundo. Un reto especialmente difícil es el amplio ámbito de la Ecología y la enorme variabilidad en las escalas de trabajo (Michener y Jones, 2012).

Los dos párrafos anteriores resumen la situación de la Ecología como ciencia computacional y como ciencia intensiva en datos. Sin embargo, si sustituyéramos Ecología por Geografía Física, la práctica totalidad de los geógrafos físicos aceptarían estas ideas.

En Ciencias de la Tierra ha aparecido una gran cantidad de términos vinculados a la incorporación de nuevas tecnologías. Tenemos la *Geomática* (Gomarasca, 2009), la *Geoinformática*

[1] *Data-intensive science*

(Karimi, 2009), la *Ciencia de la Información Geográfica* (Kemp, 2008) o la *Geocomputación* (Longley et al., 1998; Abrahart et al., 2000). La distinción entre unos y otros conceptos resulta en ocasiones difusa.

La *Geoinformática* parece ser un término paraguas que agrupa el conjunto de tecnologías recientes, sofisticadas y de muy diverso tipo que pueden utilizarse para la captura, gestión, almacenamiento y análisis de información geoespacial.

La *Geomática* se relaciona con la producción de datos mediante estas tecnológicas y se está desarrollando en el ámbito académico (Ingeniería geomática, másteres en Geomática, etc.). Gomarasca (2009) la define como un enfoque sistémico, multidisciplinar e integrado para la selección de instrumentos y técnicas apropiadas para la recolección, almacenamiento, integración, modelado, análisis, recuperación, transformación, visualización y distribución de datos georreferenciados. Estos datos, obtenidos a partir de diferentes fuentes, se caracterizan por una exactitud bien definida, continuidad y formato digital.

La *Geocomputación* (Longley et al., 1998; Abrahart et al., 2000) hace referencia a una Geografía computacional, mientras que la *Ciencia de la Información Geográfica* es el análisis y la profundización en los conceptos, técnicas y herramientas implementados en los Sistemas de Información Geográfica. Es la "ciencia tras el sistema informático" (Kemp, 2008). *Geocomputación* y *Ciencia de la Información Geográfica* son los conceptos más ligados a la Geografía, aunque posiblemente más a la Geografía humana que a la física.

Es evidente la necesidad de la Geografía de no descolgarse de estas líneas de trabajo, aunque puedan incomodar a muchos geógrafos (Gould, 2000). Estas permiten plantear y resolver ciertos problemas como la descripción y la síntesis regional o el análisis integrado de múltiples variables físicas o humanas que antes no podían serlo (Gould, 1987).

La situación no es nueva, ya en 1957 Cholley constataba como "La cartografía representa, en la actualidad, la única rama de la geografía francesa que abre salidas profesionales a los geógrafos jóvenes, aparte de la enseñanza", (citado por Philipponneau, 2001). En la actualidad son las nuevas técnicas asociadas a la informática (bases de datos, cartografía automática, teledetección, SIG) las que crean nuevos empleos y facilitan la investigación aplicada (Philipponneau, 2001).

Unwin (1992) señalaba, y es algo comprobable aún veinte años después, que la mayoría de los empleadores que valoraban la formación en Geografía como algo útil, destacaban la importancia de los conocimientos en informática y en estadística que forman parte del programa de estudios. Eso, lógicamente, cuando esos conocimientos existen.

Desde siempre, el geógrafo de cualquier especialidad ha considerado el mapa como una herramienta esencial. En un equipo multidisciplinar, el geógrafo se considera el hombre de los mapas. Con el uso del ordenador la cartografía automática significa una revolución que le permite realizar mapas de calidad (Philipponneau, 2001). En definitiva la cartografía tradicional permitía, al igual que algunas de las nuevas técnicas geoinformáticas, almacenar, mostrar y analizar información sobre el territorio.

A pesar de una valoración positiva hacia el uso de la informática, sigue existiendo el temor a la dispersión que la especialización pudiera traer a la Geografía (Gould, 1987; Unwin, 1992; Philipponneau, 2001). Sin embargo Gregory (2000) afirma que este temor no se ha cumplido, al menos en la Geografía física, debido a que el manejo de modelos y aproximaciones conceptuales similares han establecido nuevos lazos entre las diferentes ramas de la Geografía y, por otro lado, han permitido tender puentes con otras disciplinas. En todo caso, la división en diferentes ramas es un proceso normal que ocurre en todas las ciencias (Sala y Batalla, 1996).

En realidad la Geografía física siempre ha tratado de desarrollar relaciones transfronterizas entre sus diversas ramas (Stoddart, 1986; Slaymaker y Spencer, 1998) y ha reconocido la decisiva influencia de los avances tecnológicos en su desarrollo (Gardner, 1996).

Slaymaker y Spencer (1998) consideran que la Geografía física se ha convertido en una agrupación de ciencias naturales con una fuerte dependencia en el desarrollo y aplicación de técnicas de control, análisis y modelización. Para estos autores, un mayor y mejor uso de herramientas como los SIG y la teledetección pueden ayudar a superar el enfoque, excesivamente reduccionista según él, habitual en Geografía física, y recuperar una visión holística consciente de la variedad de efectos de retroalimentación (Slaymaker y Spencer, 1998). Hoy en día estas tendencias se asumen con naturalidad, como puede comprobarse en manuales recientes de Geografía física (Holden, 2008) o en los artículos publicados en *Progress in Physical Geography*.

3. CÓDIGO ABIERTO Y DISPONIBILIDAD DE DATOS

Tanto para la ciencia computacional como para la ciencia intensiva en datos, el ordenador pasa de ser una herramienta auxiliar a ser el auténtico protagonista de la investigación. Las nuevas técnicas de análisis de datos incluyen la modelización (Wainwright y Mulligan, 2004) o técnicas estadísticas sofisticadas que, aprovechando las actuales capacidades de cómputo masivo, dejan de apoyarse en métodos paramétricos. Estos asumen que el análisis de unos pocos estadísticos, representativos de los datos, bastan para obtener conclusiones. Con el nuevo enfoque, basado en modelos estadísticos que superan la regresión lineal (Zuur et al., 2009) o en métodos de aprendizaje automático (Kanevski et al., 2009), las conclusiones se obtienen con un análisis del conjunto de los datos.

El científico deja de utilizar los algoritmos cerrados que acompañan a un determinado programa y comienza a crear sus propios algoritmos. Estos pueden ser listados más o menos estructurados de órdenes en programas modulares como GRASS o R (*scripts*) o programas escritos mediante lenguajes de alto nivel que acceden a librerías científicas especializadas. Estos algoritmos, y el código informático que los materializa, pasan a ser un elemento clave en la investigación.

Una de las ventajas más habitualmente mencionadas del software abierto es precisamente su modularidad, cualidad que permite modificarlo para adaptarlo a las necesidades específicas de cada proyecto o investigación (Câmara et al., 2012).

Un primer problema que se plantea es que los sistemas interactivos de software no permiten a los usuarios mantener un seguimiento de las acciones realizadas, no permiten un registro automático de los diferentes pasos de ejecución de una determinada metodología o algoritmo (Mesirov, 2010; Peng, 2011). Por otro lado no permiten automatizar los procesos, lo cual resulta muy útil cuando se manejan grandes volúmenes de datos.

Mesirov (2010) afirma que este uso cada vez mayor de la computación plantea nuevos desafíos para la publicación científica y la replicabilidad de la ciencia. Grandes conjuntos de datos se analizan y reanalizan, modificando métodos y parámetros, y a veces, incluso actualizando los datos, hasta obtener los resultados finales, aun así, la publicación resultante suele prestar poca atención a los detalles de cálculo.

Se ha llegado a sugerir que un artículo acerca de resultados computacionales es publicidad, no conocimiento, estando el auténtico conocimiento en el entorno de software, código y datos empleados (Claerbout y Karrenbach, 1992, citado por Donoho, 2010). Sin embargo, en muchas ocasiones, el código o conjunto de códigos que dieron origen a los resultados pueden haberse perdido o ser irrecuperables.

McCafferty (2010) e Ince et al. (2012) sintetizan una serie de opiniones apoyando la necesidad de liberar el código generado y utilizado, y la necesidad de exigirlo de cara a la publicación de un trabajo científico. Por una parte, está generalmente pagado con dinero público; por otra, es necesario analizar la calidad del código, ya que de ella depende la calidad de los resultados científicos.

Los errores en el código son ubicuos en la ciencia computacional y estos pueden alterar la interpretación de los resultados. A menudo suele olvidarse lo que se ha hecho, dando lugar a interpretaciones erróneas de los resultados (Donoho, 2010). Si se describe el código en una publicación, la ambigüedad de esta descripción supone, además, que cualquier intento de reproducir dicho código puede dar lugar a programas diferentes con resultados también diferentes (Ince et al., 2012).

Las preocupaciones por la importancia del código han llevado también a mirar con cierto recelo el código propietario. El software propietario no está libre de errores, pudiendo estar mal documentado internamente y poco probado (Barnes, 2010). Como alternativa, el software de código abierto, independientemente de su calidad original, acaba mejorando por las contribuciones de la comunidad científica y permite a otros compartir la línea de investigación del autor. Rocchini y Neteler (2012) plantean la necesidad de adoptar la filosofía del software libre en Ecología. En el campo de los SIG se han hecho diversas propuestas en el mismo sentido (Ramsey, 2007; Jolma et al., 2008; Hall y Lehay, 2008; Steiniger y Bocher, 2009). El gran éxito de un programa de análisis estadístico de código abierto como R (Vance, 2009; Smith, 2010) refleja la necesidad que la sociedad tienen de este tipo de software.

Steiniger y Bocher (2009) señalan como una de las grandes ventajas del software abierto, la independencia del usuario respecto a cambios en la política de licencias por parte de la empresa propietaria del código, así como los cambios en los lenguajes de programación soportados. Estos vambios han generado bastantes problemas a los usuarios de software comercial para SIG. Por otro lado, estos mismos autores señalan que la publicación de nuevas versiones del software comercial, al tener una motivación puramente económica, puede retrasar la adopción de nuevos modelos y algoritmos.

La liberación del código implica, además de beneficios obvios para la sociedad, beneficios también para el investigador. En primer lugar mejora los hábitos de trabajo y el trabajo en equipo, y por otro lado permite un mayor impacto ya que otros pueden utilizar el código y citarlo, garantizando la continuidad de las líneas de trabajo (Donoho, 2010).

Recientemente han aparecido en las revistas científicas de referencia una serie de artículos preocupados por el problema de la repetibilidad de los estudios científicos cuando estos implican el uso intensivo de medios informáticos. Algunas revistas de referencia (*Science, Biostatistics, Geoscientific Model Development*) están empezando a exigir a los autores la entrega del código con el que se han obtenido los resultados presentados (Ince et al., 2012).

Similares preocupaciones se han producido en relación a los datos. Vision (2010) afirma que el compartir datos incrementa el valor de estos, ya que permitiría a otros estudiar y mejorar métodos de análisis, así como reinterpretar resultados indefinidamente. Este autor llama así a la reusabilidad de los datos científicos y su liberación al dominio público. Los datos no liberados de este modo correrían el riesgo de perderse, corromperse o quedar obsoleto su formato, con lo que serían irrecuperables.

Las nuevas tecnologías han incrementado la capacidad de recogida de datos, la cantidad de datos disponibles y la posibilidad de que otros investigadores los reanalicen en un proceso global de minería de datos. Poner los datos a disposición de la comunidad científica es ya un elemento esencial en la investigación. *Science* y otras revistas tienden a publicarse *online* añadiendo material de apoyo (Hanson et al., 2011).

Para resolver el problema del acceso abierto a datos y programas se plantean retos sociológicos y tecnológicos (Reichman et al., 2011). Respecto a los primeros, estos autores apuntan a la posibilidad de que la financiación de proyectos y la publicación de resultados estuviera sujeta a que los datos estén a disposición de la comunidad. En cuanto a los retos tecnológicos, plantean tres:

- Dispersión geográfica.
- Heterogeneidad debida a la diversidad de subdisciplinas dentro de la Ecología, es necesario un avance en ontologías.
- Historia[2] de los datos, entendiendo por tal los flujos de trabajo y transformación que han experimentado.

Estos problemas han sido ya planteados, y en parte resueltos, para los datos espaciales mediante las Infraestructuras de Datos Espaciales. Una IDE es una iniciativa que pretende crear un entorno en el que todas las partes pueden cooperar entre sí e interactuar con la tecnología, para lograr mejor sus objetivos en los diferentes niveles políticos y administrativos (Williamson, 2003). El objetivo es solventar el problema de la interoperabilidad (múltiples formatos de datos, muchos de ellos cerrados) y la dificultad de localizar la información (Granell et al., 2009). En la década de los 80 y 90 del siglo XX, estos problemas dificultaban considerablemente la labor de diferentes administraciones y la gestión territorial, afectando negativamente a la economía.

La directiva INSPIRE establece que "los elementos componentes de las infraestructuras deberán incluir metadatos, conjuntos de datos espaciales y servicios de datos espaciales, servicios y tecnologías de red, los acuerdos sobre el intercambio, acceso y uso; así como mecanismos, procesos y procedimientos de coordinación y supervisión" (Commission of The European Communities, 2004). Sin embargo en una IDE los datos no son realmente libres, tienen diversas licencias. Un proyecto de datos espaciales realmente abiertos es *OpenStreetMaps* (Steiniger y Bocher, 2009).

4. PROPUESTA DE UN SIG PARA EL MANEJO INTENSIVO DE DATOS

En esta sección se propone un entorno computacional para trabajar con datos espaciales. Podría considerarse un Sistema de Información Geográfica en sentido amplio. A partir de las consideraciones anteriores podemos afirmar que este sistema debe basarse en software abierto, modular y que permita la programación.

Una de las características fundamentales del software abierto es que los programas que pueden trabajar de forma conjunta, se desarrollan para conseguir la mejor coordinación posible. Se evita así también reinventar la rueda. En GRASS, por ejemplo, se abandonó el desarrollo de herramientas de análisis estadístico en el momento que fue evidente que era preferible caminar hacia una mayor integración con R.

Câmara et al. (2012) dividen los programas de SIG de código abierto en 5 grupos:

1. Programas de visualización y análisis basados en Java.
2. Programas de visualización y análisis basados en C/C++.
3. Paquetes de análisis raster.
4. Interfaces con gestores de bases de datos.
5. Librerías de apoyo.

El sistema propuesto incluye elementos de los 4 últimos grupos, se prevé un sistema de visualización basado en C/C++ por la mayor eficiencia de este lenguaje respecto a Java. Todos los programa propuestos pueden ser implementados tanto en Windows como en GNU/Linux; y todos, salvo GMT, tiene formato nativo en Windows. Sin embargo es en un sistema GNU/Linux

[2] *Provenance* en el original

donde mejor pueden trabajar de forma coordinada. En Sherman (2008) se describen con más detalle los diferentes modos de integración de los programas aquí propuestos.

4.1. GRASS

El programa GRASS[3] (Neteler y Mitasova, 2008; Neteler et al., 2008; Neteler et al., 2012) es un SIG utilizado para la gestión, análisis y visualización de información geoespacial, procesamiento de imágenes y modelización espacial. Hoy en día es el mayor proyecto de SIG desarrollado como software libre y uno de los mayores por el tamaño del código.

Como SIG multipropósito incluye 425 módulos (Neteler et al., 2012) que le permiten llevar a cabo la mayor parte de las funcionalidades de un SIG. Cada uno de estos módulos se diseña para ser lo más robusto y eficiente posible al ejecutar una determinada tarea. El programa utiliza la filosofía Unix de encadenamiento de múltiples pequeños programas para realizar tareas complicadas. Cualquier análisis complejo suele necesitar la integración de múltiples módulos mediante *tuberías*[4] o mediante ficheros intermedios.

Tradicionalmente GRASS ha sido un programa orientado al manejo de datos ráster. Sin embargo, la rama de desarrollo 5.7, iniciada a comienzos del presente siglo, supone una modificación de gran calado, especialmente en lo que respecta a la gestión de datos vectoriales. De hecho, el final de esta rama es la actual versión 6 del programa.

Utiliza librerías externas para algunas de las tareas rutinarias en un SIG: GDAL para la importación/exportación de diversos formatos de datos geoespaciales; Proj4 para convertir coordenadas geográficas a cartesianas, así como la transformación inversa, pudiendo utilizar un amplio conjunto de proyecciones cartográficas y GEOS, una librería de funciones geométricas adaptación a lenguaje C++ de la librería de Java JTS.

Permite utilizar computación paralela en sistemas multiprocesador (Neteler, 2005, 2010).

GRASS se integra bien con otras herramientas tanto de código abierto como propietario (Sherman, 2008; Neteler et al., 2008; Neteler et al., 2012)., las expuestas a continuación son un buen ejemplo. En cada caso se apunta brevemente como se produce esta integración.

4.2. PostgreSQL

El modelo geo-relacional de bases de datos, tradicional en los SIG, implica el almacenamiento de las tablas enlazadas a las capas vectoriales en una base de datos relacional. PostgreSQL[5] es un Sistema de Gestión de Bases de Datos altamente fiable, escalable y compatible con SQL. Incluye las características habituales de las bases de datos relacionales y añade numerosas características avanzadas, incluyendo interfaces de programación para diversos lenguajes (Chen y Xie, 2008; Matthew y Stones, 2005).

Por defecto GRASS almacena las tablas enlazadas a un mapa vectorial en formato DBF. Este comportamiento puede modificarse para utilizar otros gestores de bases de datos (por ejemplo PostgreSQL). En GRASS existen diversos módulos para gestionar el enlace de tablas y mapas o para la visualización y consulta de estas tablas. Estos módulos actúan, por tanto, como programas clientes de bases de datos.

---

[3] http://grass.fbk.eu/

[4] Herramienta de los sistemas UNIX para encadenar dos programas de manera que la salida de datos del primero sea la entrada de datos del segundo. De esta manera se eliminan los ficheros intermedios y se facilita la automatización de procesos.

[5] http://www.postgresql.org/

Existe una gran diferencia entre utilizar ficheros DBF o tablas en una base de datos PostgreSQL. En el primer caso apenas puede recuperarse poco más que el valor de una columna mientras que trabajando con PostgreSQL pueden lanzarse consultas complejas en SQL, y el servidor de bases de datos responderá de forma eficiente. Los módulos para la visualización de capas vectoriales actúan del mismo modo, permitiendo seleccionar aquellos objetos dentro de una determinada capa que pueden visualizarse en cada momento.

PostgreSQL permite agrupar las tablas en una base de datos en esquemas de usuario, de manera que cada usuario dispone de su propio espacio de trabajo. El concepto Base de datos-Esquema-Tabla de PostgreSQL es equivalente al concepto LOCATION-MAPSET-Capa de GRASS. Resulta muy conveniente entonces que todas las tablas de una LOCATION se almacenen en una base de datos y que haya una correspondencia biunívoca entre MAPSETS y esquemas.

La gran limitación de SQL en relación con los SIG es su incapacidad para procesar consultas espaciales. Una solución desarrollada en los últimos años es la utilización de bases de datos objeto-relacionales como punto de partida para crear bases de datos espaciales (Manolopoulos et al., 2004). Este nuevo modelo admite tipos de datos abstractos que incluyen objetos geométricos, al mismo tiempo se define un SQL extendido, lo que permite funciones y operadores espaciales e índices espaciales para acelerar las consultas y operaciones.

PostGIS[6] es una extensión de PostgreSQL que añade tipos de datos, operadores y funciones para el manejo de objetos geoespaciales, convirtiendo a este programa en un gestor de bases de datos espaciales. Cumple la OpenGIS "Simple Features Specification for SQL" (Sherman, 2008; Obe y Hsu, 2011). GRASS puede importar y exportar datos de PostGIS utilizando la librería GDAL.

A pesar de que las últimas versiones de GRASS admiten datos vectoriales, los formatos no cumplen los estándares de la OGC ni las herramientas de geoprocesamiento son tan potentes como en el caso de PostGIS.

4.3. QGIS (Quantum GIS)

Desde el principio GRASS se concibió como un programa dirigido al análisis y modelización de datos espaciales, y no como un programa de cartografía automática o *Desktop mapping*; por ello su interfaz gráfica se concibió como un medio (austero, robusto y totalmente portable) y no como un fin en sí mismo. La falta de una Interfaz Gráfica de Usuario definida es posiblemente uno de los tradicionales inconvenientes a la hora de empezar a trabajar con él.

Se han desarrollado diversas interfaces gráficas para el programa utilizando diversos lenguajes de programación. Sin embargo, hemos preferido utilizar QGIS[7] para esta tarea. Se trata de un SIG de escritorio que permite leer información en múltiples formatos (incluyendo GRASS y PostGIS) así como establecer conexiones WMS.

La idea original del equipo de desarrollo fue proveer a la comunidad de usuarios de SIG libre de un visor de datos geográficos rápido y sencillo para ordenadores con sistema operativo GNU/Linux. Aunque en la actualidad se puede asumir que se ha superado con creces este objetivo, ya que incorpora numerosas funcionalidades de análisis de datos espaciales.

Una de las más importantes es la posibilidad de comportarse como un visor de datos y una interfaz gráfica de GRASS (Sherman, 2008; Steiniger y Bocher, 2009). El principal problema para manejar QGIS con GRASS en un entorno multiusuario es que para leer datos de GRASS, QGIS obliga a que el usuario sea el propietario de la LOCATION (directorio de trabajo de GRASS).

---

[6] http://postgis.refractions.net/

[7] http://www.qgis.org/

De esta manera, dos usuarios no podrían trabajar al mismo tiempo en la misma LOCATION, lo que entra en contradicción con el espíritu multiusuario de GRASS.

Para solventar este problema se puede trabajar de manera que cada usuario sea propietario de su LOCATION, pero que bajo ella aparezcan enlazados los MAPSETS (conjuntos de mapas) comunes. En Neteler y Mitasova (2008) aparece una explicación más detallada de la estructura de directorios en GRASS.

4.4. R

R[8] es ante todo una herramienta de modelización estadística que incorpora utilidades gráficas de análisis exploratorio y un eficiente lenguaje de programación orientado a objetos (Crawley, 2005). El lenguaje R permite al usuario, por ejemplo, programar bucles para analizar conjuntos sucesivos de datos. También es posible combinar en un solo programa diferentes funciones estadísticas para realizar análisis más complejos.

El uso de las funciones es relativamente intuitivo, los argumentos pueden ser objetos de cualquier tipo (datos, fórmulas, expresiones), algunos de los cuales se definen por defecto. Una descripción detallada de R está disponible en Venables et al. (2012).

El programa crece fácilmente merced a la comunidad de usuarios que programan paquetes. Gran parte de estos han sido creados con el lenguaje de programación de R. En este momento hay disponibles 3.848 paquetes dedicados a diferentes ciencias y métodos de análisis de datos, de ellos más de cien permiten trabajar con datos espaciales.

Entre ellos cabe destacar *spgrass6* que permite leer y escribir capas en formato de GRASS (Bivand, 2007) desde R. De manera que cualquier capa de información espacial almacenada en formato de GRASS puede ser leída y analizada con R.

Otros paquetes de interés para la Geografía física son, a modo de ejemplo:

- *sp*, incluye clases y métodos para el manejo de datos espaciales. Surgió de un esfuerzo de integración de diferentes paquetes anteriores (Pebesma y Bivand, 2005).
- *rgdal*, permite acceder a las librerías gdal y proj4 (Keitt et al., 2010).
- *topmodel*, implementación del modelo hidrológico TOPMODEL (Buytaert, 2011).
- *landsat*, incluye funciones para la corrección de imágenes de satélite (Goslee, 2011).
- *gstat*, incluye la gran mayoría de las funciones del progrma de análisis geoestadístico *gstat* (Pebesma, 2004).

La gran ventaja de la integración de R con GRASS es que todas las acciones en R se realizan con objetos guardados en la memoria activa del ordenador, sin usar archivos temporales, mientras que GRASS almacena toda la información en ficheros y carga muy poco la memoria. De esta manera, ambos programas se reparten los recursos sin cargar el sistema. Por ejemplo, puede leerse una capa de puntos, generar una capa interpolada con el paquete *gstat* y guardar los resultados como fichero de GRASS.

4.5. GMT

GMT[9] pone a disposición del usuario un conjunto de módulos orientados a la producción de cartografía (Sherman, 2008; Wessel y Smith, 2012). El manejo de estos módulos en línea de comandos y la posibilidad de combinarlos (entre sí y con otras herramientas UNIX), así como el elevado número de opciones de cada uno de ellos, convierte GMT en un entorno de maquetación

[8] http://cran.r-project.org/
[9] http://gmt.soest.hawaii.edu

de mapas extremadamente flexible. La contrapartida de esta flexibilidad es un lenguaje que puede llegar a ser bastante complejo y hermético.

GMT genera mapas en formato postscript (extensión .ps) mediante la inclusión progresiva de diversos elementos al mapa/fichero por parte de diferentes módulos. Los ficheros postscript pueden transformase de forma sencilla a formato PDF.

A pesar de que utiliza formatos propios no estándar, puede integrarse con GRASS con cierta facilidad. Para ello es necesario redirigir la salida de los módulos de GRASS que se van a utilizar para visualizar en el mapa los diferentes objetos. En lugar de dibujar sobre el monitor deben hacerlo sobre un fichero gráfico (formato PNG). Este fichero se incluye en el archivo postscript de GMT y posteriormente se añaden los diferentes elementos auxiliares del mapa.

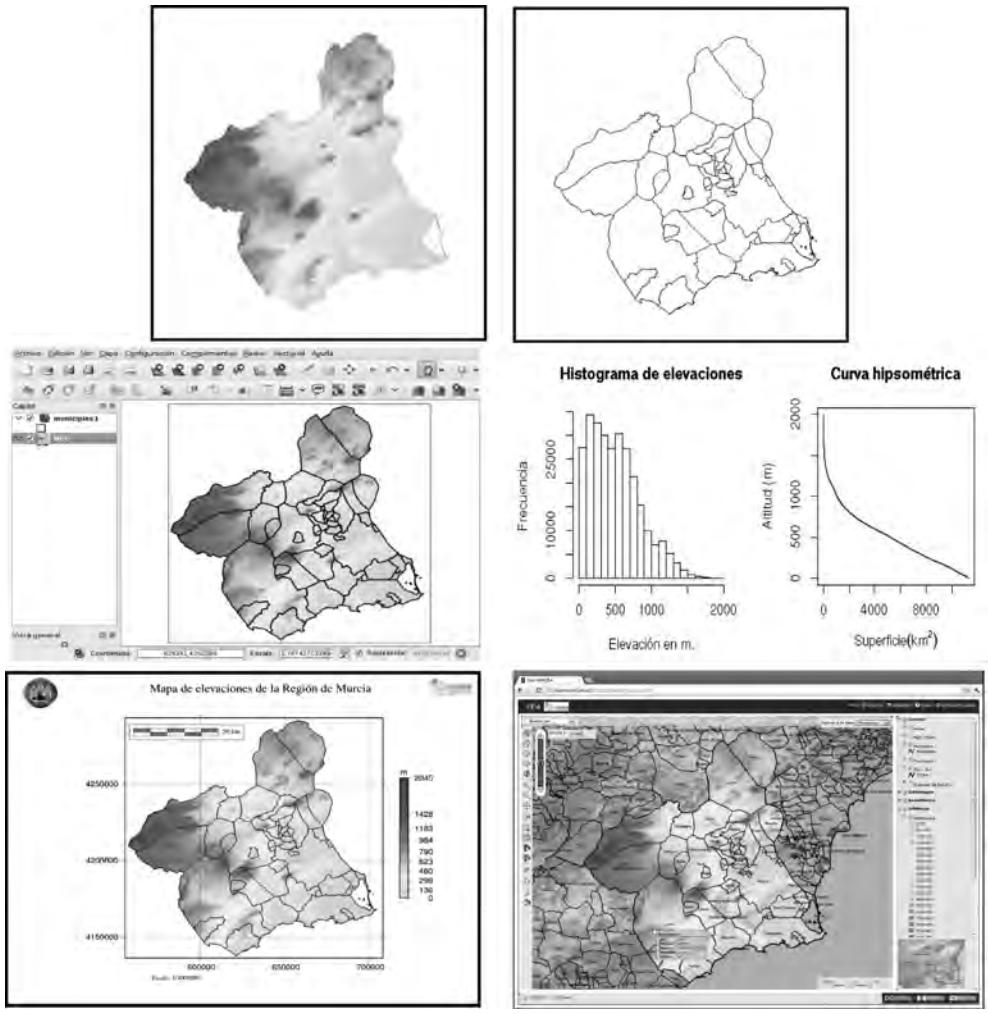

Figura 1. Integración de capas raster de GRASS y vectorial de PostGIS (arriba). Visualización con QGIS y análisis estadístico con R (centro), mapa maquetado con GMT (abajo izquierda) y capas servidas por WMS con MapServer (abajo derecha).

4.2. MapServer

MapServer[10] (Lime, 2008) es una plataforma para construir aplicaciones y servicios de Internet capaces de manejar datos espaciales. Permite generar de forma automática los diversos objetos que van a formar parte de un mapa interactivo en el que los modos de visualización pueden hacerse dependientes de la escala.

Incluye una aplicación CGI que permite construir fácilmente sitios web interactivos para mostrar información espacial y una interfaz para acceder a la API de MapServer escrita en C. Esta interfaz permite desarrollar aplicaciones más complejas con diversos lenguajes de scripts (PHP, Perl, Python) e incluso con Java.

El elemento básico para crear una aplicación de MapServer es el fichero *mapfile*. Se trata de un fichero de texto que, de manera jerárquica, incluye los diferentes elementos que pueden formar parte del mapa. Cada capa de información espacial se introduce como un elemento independiente, incluyendo la ubicación de los datos (fichero o base de datos); qué intervalos de valores se van a representar y, en el caso de capas vectoriales, en qué columna de la taba aparecen; qué paleta de color se va a utilizar; y qué etiquetas mostrar. Permite también definir estilos de tramas, iconos para representar puntos y los modos en que la información temática se consulta y muestra al usuario.

5. CONCLUSIÓN

En este trabajo se han hecho una serie de consideraciones acerca de la importancia creciente de la computación en las ciencias ambientales en general y en la Geografía física en particular. Se ha destacado la importancia de utilizar software abierto que incluya herramientas adecuadas de programación de tareas que permitan un escrutinio a posteriori del trabajo realizado, así como la replicabilidad del mismo. Finalmente se ha propuesto un entorno de trabajo basado en software abierto, con herramientas perfectamente integrables y programables. Este entorno se basa en GRASS, PostgreSQL/PostGIS, R, QGIS, GMT y Mapserver.

REFERENCIAS

ABRAHART, R.J., OPENSHAW, S. y SEE, L.M. (Eds.) (2000): *Geocomputation*. CRC Press.

BARNES, N. (2010): Publish your computer code: it is good enough, *Nature*, 467, pp. 753.

BELL, G., HEY, T. y SZALAY, A. (2009): Beyond the Data Deluge, *Science*, 323 (5919), pp. 1297-1298.

BIVAND, R. (2007): Úsing the R-GRASS Interface: Current Status, *OSGeo Journal*, 1, pp. 36-38.

BOETTINGER, J.L., HOWELL, D.W., MOORE, A.C., HARTEMINK, A. E. y KIENAST-BROWN, S. (Eds.) (2010): *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation*, Springer.

BUYTAERT, W. (2011): *Topmodel: Implementation of the hydrological model TOPMODEL in R*. R package version 0.7.2-2. http://CRAN.R-project.org/package=topmodel

[10] http://mapserver.org/

CÂMARA, G., VINHAS, L. y DE SOUZA, R.C.M. (2010): Free and Open Source GIS: Will there ever be a Geo-Linux?. En *Geospatial Free and Open Source Software in the 21st Century*, editado por E. Bocher y M. Neteler, Springer, Berlin, pp. 229-246.

CHEN, R. y XIE, J. (2008): Open Source Databases and Their Spatial Extensions. En *Open Source Approaches in Spatial Data Handling*, editado por B. Halll y M.G. Leahy, Springer, pp. 105-130.

COMMISSION OF THE EUROPEAN COMMUNITIES (2004): *Proposal for a Directive of the European Parlament and of the Council establishing an infrastructure for spatial information in the Community (INSPIRE)*. http://inspire.jrc.it/proposal/EN.pdf

CRAWLEY, M.J. (2005): *Statistics: An Introduction using R*. Willey.

DONOHO, D.L. (2010): An invitation to reproducible computational research, *Biostatistics*, 11(3), pp. 385-388.

GARDNER, R. (1996): Developments in Physical Geography En: E.M. Rawling y R.A. Daugherty (Eds.), Geography into the twenty-first century, pp. 95112. Willey.

GOMARASCA, M. A. (2009): *Basics of Geomatics*. Springer.

GOSLEE, S.C. (2011): Analyzing Remote Sensing Data in R: The landsat Package, *Journal of Statistical Software*, 43(4), pp. 1-25. http://www.jstatsoft.org/v43/i04/

GOUDIE, A. (1994): The Nature of Physical Geography: A View from the Drylands Geography, 79 (3), pp. 194209.

GOULD, P. (1987): Pensamientos sobre la Geografía, *Geocrítica*, 68.

GOULD, P. (2000): Pensar como un geógrafo. Una exploración en la Geografía moderna, *Scripta Nova*, 78.

GRANELL, C., GOULD, M., MANSO, M.A. y BERNABE, M.A. (2009): Spatial Data Infrastructures. En *Handbook of Research on Geoinformatics*, editado por H.A. Karimi, Information Science Reference, New York, pp. 36-41.

GREGORY, K.J. (2000): *The changing Nature of Physical Geography*. Arnold.

HALL, G.B. y LEHAY, M.G. (Eds.) (2008): *Open Source Approaches in Spatial Data Handling*. Springer.

HANSON, B., SUGDEN, A. y ALBERTS, B. (2011): Making Data Maximally Available, *Science*, 331, pp. 649.

HEY, T., TANSLEY, S. y TOLLE, K. (Eds.) (2009a): *The Fourth Paradigm Data-Intensive cientific Discovery*, Microsoft Research, Redmond, Whasington.

HEY, T., TANSLEY, S., y TOLLE, K. (Eds.) (2009b): *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research.

HOLDEN, J. (Ed.) (2008): *An Introduction to Physical Geography and the Environment*, Pearson Prentice Hall.

HUNT, J. R., BALDOCCHI, D.D. y VAN INGEN, C. (2009): Redefining ecological science using data. En *The Fourth Paradigm Data-Intensive Scientific Discovery*, editado por T. Hey, S. Tansley y K. Tolle, Microsoft Research, Redmond, pp. 21-26.

INCE, D.C., HATTON, L. y GRAHAM-CUMMING, J. (2012): The case for open computer programs, *Nature*, 482, pp. 485-488.

JOLMA, A., AMES, D.P., HORNING, N., MITASOVA, H., NETELER, M., RACICOT, A. y SUTTON, T. (2008): Free and Open Source Geospatial Tools for Environmental Modelling and Management. En *Environmental Modelling, Software and Decision Support*, editado por Anthony J. Jakeman, Alexey A. Voinov, Andrea E. Rizzoli y Serena H. Chen Elsevier, pp. 163-180.

KANEVSKI, M., POZDNOUKHOV, A. y TIMONIN, V. (2009): *Machine learning for spatial environmental data. Theory, applications and software*. EPFL Press.

KARIMI, H. A. (Ed.) (2009): *Handbook of Research on Geoinformatics*. Information Science Reference.

KEITT, T.H., BIVAND, R., PEBESMA, E. Y ROWLINGSON, B. (2010): *rgdal: Bindings for the Geospatial Data Abstraction Library*. R package version 0.6-33. http://CRAN.R-project.org/package=rgdal

KEMP, K-K. (Ed.) (2008): *Encyclopedia of Geographic Information Science*. SAGE publications.

KUMAR, P., ALAMEDA, J.C., BAJCSY, P., FOLK, M. y MARKUS, M. (2005): *Hydro-informatics*. Taylor & Francis.

LAGACHERIE, P., MCBRATNEY, A.B. y VOLTZ, M. (2007): *Digital Soil Mapping: An Introductory Perspective*. Elsevier.

LIME, S. (2008): MapServer. En *Open Source Approaches in Spatial Data Handling*, editado por B. Hall y M.G. Leahy, Springer, pp. 65-85.

LONGLEY, P.A., BROOKS, S.M., MCDONNELL, R. y MACMILLAN, B. (Eds.) (1998): *Geocomputation: A Primer*. Willey-Blackwell.

MANOLOPOULOS, Y., PAPADOPOULOS, A.N. y VASSILAKOPOULOS, M.Gr. (2004): *Spatial Databases: Technologies, Techniques and Trends*. Idea Group Publishing.

MATTHEW, N. y STONES, R. (2005): *Beginning Databases with PostgreSQL*. Apress.

MCCAFFERTY, D. (2010): Should Code be Released?, *Communication s of the ACM*, 55(10), pp. 16-17.

MESIROV, J.P. (2010): Accessible Reproducible Research, *Science*, 327, pp. 415-416.

MICHENER, W.K. y JONES, M.B. (2012): Ecoinformatics: supporting ecology as a data-intensive science Trends in Ecology and Evolution. *Trends in Ecology and Evolution*, 27(2), pp. 85-93.

NETELER, M., BEAUDETTE, D.E., CAVALLINI, P., LAMI, L. y CEPICKY, J. (2008) : GRASS SIG. En *Open Source Approaches in Spatial Data Handling*, , editado por B. Hall y M.G. Leahy, Springer, pp. 171-199.

NETELER, M., BOWMAN, M.H., LANDA, M. y METZ, M. (2012): GRASS GIS: a multi-purpose Open Source GIS, *Environmenal Modelling & Software*, 31, pp. 124-130.

NETELER, M. y MITASOVA, H. (2008): *Open Source GIS: A GRASS GIS Approach*. Springer.

NETELER, M. (2005): Time Series Processing of MODIS Satellite Data for Landscape Epidemiological Applications, *International Journal of Geoinformatics*, 1(1), pp. 33-137.

NETELER, M. (2010): Estimating Daily Land Surface Temperatures in Mountainous Environments by Reconstructed MODIS LST Data, *Remote Sensing*, 2, pp. 333-351.

OBE, R.O. y HSU, L.S. (2011): *PostGIS in Action*. Manning Publications.

PEBESMA, E.J. (2004): Multivariable geostatistics in S: the gstat package, *Computers & Geosciences*, 30, pp. 683-691.

PEBESMA, E.J. y BIVAND, R.S. (2005): Classes and methods for spatial data in R. *R News*, 5 (2), 9-13. http://cran.r-project.org/doc/Rnews

PENG, R.D. (2011): Reproducible Research in Computational Science, *Science*, 334, pp. 1226-1227.

PHILIPPONNEAU, M. (2001): *Geografía aplicada.* Ariel Geografía.

RAMSEY, P. (2007): The state of open source GIS. En *FOSS4G 2007 conference*.

REICHMAN, O. J., JONES, M.B. y SCHILDHAUER, M.P. (2011): Challenges and Opportunities of Open Data in Ecology, *Science*, 331, pp. 703-705.

ROCCHINI, D. y NETELER, M. (2012): Let the four freedoms paradigm apply to ecology, *Trends in Ecology and Evolution*,, 27(8), pp. 310-311.

SALA, M. y BATALLA, R. (1996): *Métodos y técnicas en Geografía Física*. Ed. Síntesis.

SHERMAN, G.E. (2008): *Desktop GIS. Mapping the Planet with Open Source Tools*. The Pragmatic Programmers.

SLAYMAKER, O. y SPENCER, T. (1998): *Physical Geography and Global Environmental Change*. Adison Wesley.

SMITH, D. (2010): *R is Hot. How Did a Statistical Programming Language Ingvented in New Zealand Become a Global Sensation*. Informe técnico, Revolution Analytics.

STEINIGER, S. y BOCHER, E. (2009): An overview on current free and open source desktop GIS developments, *International Journal of Geographical Information Science*, 23(10), pp. 1345-1370.

STODDART, D.R. (1986): *On Geography and Its History*. Blackwell.

TRICART, J. y KILIAN, J. (1979): *L'eco-Géeographie et l'aménagement du milieu naturel*. Maspero.

UNWIN, T. (1992): *El lugar de la Geografía*. Cátedra.

VANCE, A. (2009): Data Analysts Captivated by R's Powerfi, *The New York Times*, pp. 7-1.

VENABLES, W. N., SMITH, D. M. y R DEVELOPMENT CORE TEAM (2012): An Introduction to R. En *R notes R: A Programming Environment for Data Analysis and Graphics*, version 2.15.0 (2012-03-30). http://cran.r-project.org/doc/manuals/R-intro.pdf

VISION, T.J. (2010): Open Data and the Social Contract of Scientific Publishing, *BioScience*, 60(5), pp. 330-331.

WAINWRIGHT, J. y MULLIGAN, M. (Eds.) (2004): *Environmental Modelling: Finding Simplicity in Complexity.* Willey.

WESSEL, P. Y SMITH, W.H.F. (2012): *The Generic Mapping Tools GMT. Technical Reference and Cookbook*. http://gmt.soest.hawaii.edu/gmt/pdf/GMTDocs:pdf

WILLIAMSON, I. (2003): SDIs-Setting the Scene. En *Developing Spatial Data Infrastructures: From Concept to Reality*, editado por Ian P. Williamson, Abbas Rajabifard y Mary-Ellen F. Feeney, Taylor & Francis, pp. 1-16.

ZUUR, A.F., IENO, E.N., WALKER, N., SAVELIEV, A.A. y SMITH, G.M. (2009): *Mixed Effects Models and Extensions in Ecology with R.* Springer.