Universidad de Granada
Departamento de Lenguajes y Sistemas Informáticos

Tesis doctoral

Programa Oficial de Doctorado en Tecnologías de la Información y la Comunicación

---

Nuevos métodos para el procesamiento y análisis de información geográfica

Romel Vázquez Rodríguez

Director:
Prof. Dr. Juan Carlos Torres Cantero

Granada, 2015

Editor: Universidad de Granada.Tesis Doctorales
Autor: Romel Vázquez Rodríguez
ISBN: 978-84-9125-278-8
URI: http://hdl.handle.net/10481/41303

El doctorando Romel Vázquez Rodríguez y el director de la tesis Juan Carlos Torres Cantero garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección del director de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Granada, 2 de julio de 2015

Director de la Tesis
Fdo.: Juan Carlos Torres Cantero
———————————-

Doctorando
Fdo.: Romel Vázquez Rodríguez
———————————-

La gota abre la piedra, no por su fuerza sino por su constancia

— Publio Ovidio

Dedico esta tesis a la lucha contra el cáncer.

Agradecimientos

Me gustaría agradecer a las siguientes personas e instituciones por su ayuda durante el desarrollo de este trabajo:

- En primer lugar a mis padres, por el apoyo incondicional durante mi vida, mi formación y por ayudarme a culminar este proyecto.
- A mi tutor Juan Carlos, sin sus consejos y recomendaciones hubiera sido imposible llegar hasta aquí.
- A mi cotutor cubano Carlos Perez Risquet, a él támbien le debo haber llegado hasta aquí.
- A mis tesiantes, fue un placer haber dirigido sus tesis de grado, que aportaron mucho a este proyecto.
- A todos los miembros del departamento de Ciencias de la Computación de la Universidad Central ”Marta Abreu”de Las Villas, quienes aportaron ideas que complementaron este trabajo.
- A todos los profesores que han tenido que ver con mi formación profesional en todas las etapas de mi vida.
- A todos los profesores de doctorado en Softcomputing de la Universidad de Granada y del curso de experto en sistemas de información geográfica.
- A todos los que han estado pendiente de esta investigación.
- A todos los que de una manera u otra hicieron más placentera mis estancias en España, lejos de mi familia.
- A Yanet, por soportarme hablando de estos temas durante los peores momentos de su enfermedad.
- A mi hijo, por ser mi motor impulsor para llevar adelante este proyecto.
- A todos los miembros de mi familia que siempre estuvieron pendientes de mi formación
- A la Asociación Universitaria Iberoamericana de Postgrado y a la junta de Andalucía, sin su apoyo mis estancias en España no se hubieran materializado, todas fueron importantes para la culminación exitosa de este trabajo.
- A todos los expertos que me ayudaron a valorar este trabajo.
- A todos, muchas gracias.

Índice general

Índice de tablas

Índice de figuras

Resumen

La geovisualización es un campo multidisciplinario que involucra la cartografía, la visualización científica, el análisis de imágenes, la visualización de información, el análisis exploratorio de datos y la ciencia de los sistemas de información geográfica con el objetivo de proporcionar la teoría, los métodos y las herramientas para la exploración visual, el análisis, la síntesis y la presentación de datos que contengan información geográfica.

En este trabajo se ha creado un modelo que permite integrar herramientas de visualización en sistemas de información geográfica para el análisis de grandes volúmenes de datos espacio-temporales, brindando la posibilidad de analizar simultáneamente múltiples variables geo-referenciadas, que pueden representar series temporales. Este modelo se llevó a cabo mediante el desarrollo de herramientas para el análisis exploratorio de datos con baja y alta densidad espacial, incorporando en este último caso la manipulación de formatos de datos científicos.

El análisis exploratorio de datos con baja densidad espacial fue tratado como un caso especial que permite analizar visualmente datos multiparamétricos de manera independiente o coordinada asociados a pocos lugares del espacio. Se obtuvo una herramienta con múltiples técnicas de visualización de datos multiparamétricos que fueron integradas como una extensión de la versión 1.9 de gvSIG. Se presentó un caso de estudio con datos climáticos de la provincia de Villa Clara, Cuba.

Se aplicó el modelo propuesto para el análisis exploratorio de datos con alta densidad espacial y se obtuvo una herramienta que permite la extracción de conocimiento, reconocimiento de patrones, tendencias, anomalías y relaciones entre múltiples variables distribuidas uniformemente sobre un territorio. Se propuso la integración de un conjunto de técnicas de visualización de datos multiparamétricos en la versión 1,12 de gvSIG.

Se desarrollaron un conjunto de herramientas y algoritmos que permiten la manipulación de los formatos de datos científicos HDF y netCDF en sistemas de información geográfica. Estos fueron implementados como una extensión de la biblioteca Sextante, por lo que pueden ser utilizadas en cualquier sistema de información geográfica basado en Java que permita la integración de esta biblioteca. La extensión desarrollada en Sextante es de gran utilidad para la transformación de formatos de datos comunes en los sistemas de información geográfica para formatos de datos científicos y viceversa. Además, algunos de los algoritmos están di-

señados para facilitar la creación automática de conjuntos de datos con la estructura necesaria que requiere el módulo de visualización científica de gvSIG 1,12. Se demostró el uso de las herramientas desarrolladas mediante un caso de estudio que evidencia su utilidad para crear grandes conjuntos de datos que pueden ser analizados mediante visualizaciones en sistemas de información geográfica.

En la sección 7.2 se aplicó el método de expertos para realizar una validación de las propuestas realizadas en esta tesis. Mediante los casos de estudio descritos en el trabajo y la valoración de los expertos se ha comprobado la viabilidad de la utilización del modelo para desarrollar herramientas que permiten la extracción de conocimiento, reconocimiento de patrones, tendencias, anomalías y relaciones entre múltiples variables en diferentes áreas de aplicación. En particular se obtuvo una valoración positiva sobre la contribución de la integración de técnicas de visualización de datos multiparamétricos en sistemas de información geográfica para el análisis exploratorio de grandes volúmenes de datos científicos.

Parte I

Preliminares

1 Motivación e introducción

La geovisualización es un campo emergente que se basa en la integración de muchas disciplinas, como la cartografía, la visualización científica, el análisis de imágenes, la visualización de información, el análisis exploratorio de datos y la ciencia de los sistemas de información geográfica con el objetivo de proporcionar la teoría, los métodos y las herramientas para la exploración visual, el análisis, la síntesis y la presentación de datos que contengan información geográfica (MacEachren y Kraak, 2001; Dykes *et al.*, 2005; Kraak, 2006).

El uso de sistemas de información geográfica ha llegado a ser esencial en todos los campos relacionados con la geografía y el medio ambiente, ha pasado por las aplicaciones más clásicas de cartografía, urbanismo y gestión de recursos. Actualmente se utilizan los sistemas de información geográfica para resolver problemas tan diversos como la planificación de la extinción de incendios, el análisis de riesgos ambientales o la propagación de contaminantes.

Un sistema de información geográfica integra equipamiento de cómputo, programas y bases de datos geográficas para captar y mostrar en diferentes formas la información geográfica necesaria para resolver problemas de alta complejidad en los procesos de planificación y gestión. Así mismo, se han convertido en herramientas de trabajo muy útiles para realizar consultas de manera interactiva, y procesar y analizar la información espacial. Permiten además, la edición de datos y mapas para presentar los resultados de múltiples análisis (Bolstad, 2005).

Por otra parte, la visualización científica y la minería visual de datos se han convertido en áreas de investigación de creciente interés en los últimos años, motivado fundamentalmente por el incremento constante de los volúmenes de datos generados en muchos campos de aplicación (por ejemplo, Geología, Geofísica, Meteorología, entre otros), así como por el aumento sostenido de la potencia de las interfaces gráficas modernas, las cuales permiten generar imágenes cada vez más sofisticadas.

En las últimas dos décadas se han realizado avances significativos en el establecimiento de la visualización como una herramienta de exploración de datos flexible y fácil de usar. El análisis visual de datos es un nuevo enfoque, que se beneficia de las bondades de la percepción humana para la interpretación de imágenes, así como de los métodos computacionales automáticos, lo que permite una mejor comprensión y análisis de grandes y complejos conjuntos de datos (Keim

et al., 2008b, 2010).

La visualización científica se ocupa de encontrar una representación visual apropiada para un conjunto de datos, que permita mayor efectividad en el análisis y evaluación de los mismos. Según Rhyne (1997) posibilita la transformación de los datos numéricos o simbólicos y la información en imágenes geométricas generadas por computadora. Es una metodología para interpretar, a través de una imagen en la computadora, tanto datos de mediciones como los generados por modelos computacionales (Rhyne, 1997). La investigación y el desarrollo de la visualización científica se han centrado en cuestiones relacionadas con el renderizado de gráficos en tres dimensiones, animaciones de series temporales y visualización interactiva en tiempo real (Rhyne y MacEachren, 2004).

Una aplicación de la minería visual de datos, es que se está convirtiendo aceleradamente en una herramienta de apoyo para el modelado y simulación de procesos. El reto aquí es lograr una forma de análisis efectiva, que mediante la visualización de datos de simulaciones pueda servir para la creación y validación de hipótesis, así como la investigación sobre la estructura de modelos. Esto permite una valoración de los datos antes de ejecutar costosos experimentos.

La integración de técnicas de visualización científica en sistemas de información geográfica es una idea innovadora, que combina las ventajas y fortalezas para el análisis de datos de los dos enfoques. Esta es un área de investigación interesante dentro de la geovisualización. La integración de visualizaciones de grandes volúmenes de datos espaciales y temporales facilita la comprensión de múltiples problemas que afectan a la sociedad (Andrienko *et al.*, 2003).

El mundo real no es estático, tanto componentes espaciales como no espaciales tienen que ver con el tiempo (Hogeweg, 2000). Especialistas de muchas áreas de la ciencia visualizan, consultan y analizan la información recopilada a partir del mundo real para ayudar a la toma de decisiones. Esta información existe en los dominios espacial, temporal y temático (Hogeweg, 2000). El dominio espacial es el encargado de analizar qué se mide o se encuentra en algún lugar. El dominio temporal se encarga de analizar qué se produce en algún momento o qué existe durante un cierto tiempo. El dominio temático es el que tiene que ver con el área de aplicación que se va a analizar en el espacio y el tiempo (Hogeweg, 2000).

Una de las ciencias que mejor se ajusta a la integración de los dominios espacial, temporal y temático es la meteorología. Históricamente el análisis espacio-temporal en esta área se ha realizado mediante el uso de sistemas de información geográfica, herramientas para el análisis de series temporales, paquetes estadísticos, geoestadísticos y la animación en sistemas de visualización.

1.1. Contexto

Cuando se analizan grandes volúmenes de datos en series temporales, es común encontrarse con varios problemas: el análisis de las secuencias de una variable con simples imágenes se

dificulta debido a que se requiere tener en cuenta el tiempo, y el solapamiento de todos los mapas provoca que se tengan que analizar en secuencias de imágenes o animaciones. Los métodos y herramientas actuales presentan limitaciones para el análisis espacio-temporal de múltiples variables a la vez. Tampoco existen modelos que guíen a los desarrolladores de herramientas de geovisualización en la construcción y desarrollo sistémico de herramientas que permitan realizar análisis exploratorio de grandes volúmenes de datos científicos.

Para la conceptualización de los problemas que se abordan en esta tesis el estudio se divide en cuatro partes: la primera parte, que se presenta en el capítulo 3, está relacionada con la construcción de un modelo que permita definir la integración de técnicas de visualización científica en sistemas de información geográfica para el análisis exploratorio de datos espacio-temporales.

En la segunda parte, se presenta la aplicación del modelo a la visualización de datos espacio-temporales en datos con una baja densidad espacial, pero amplios en el tiempo. Este tipo de datos puede ser encontrado en varias ramas de la ciencia, por ejemplo, datos históricos sobre estaciones meteorológicas donde se miden distintas variables, y como sucede en algunos países con pobre infraestructura tecnológica, las estaciones están muy separadas espacialmente. Otros ejemplos lo constituyen los censos tomados en regiones delimitadas por la distribución político-administrativa (municipios, provincias, países, etc.). Estas regiones generalmente no suman un número alto. El problema en este caso está dado por el hecho de que estos datos no pueden ser analizados suficientemente bien en forma numérica, por lo que surge la necesidad de visualizarlos con el objetivo de realizar comparaciones, identificar patrones, encontrar correlaciones, detectar anomalías, variabilidad y las diferentes tendencias que se pueden presentar en las variables a lo largo del tiempo, teniendo en cuenta su ubicación espacial. Una solución a este problema se presenta en el capítulo 4, mediante el desarrollo, implementación y utilización de técnicas de visualización de datos multiparamétricos integradas dentro de sistemas de información geográfica.

La tercera parte está relacionada con un problema que se deriva de la primera: el análisis de grandes volúmenes de datos espacio-temporales con una alta densidad espacial y amplios en el tiempo. Estos datos no son tan comunes de encontrar, pero en los últimos años con el desarrollo de las tecnologías de la información y las comunicaciones, de Internet y de la aeronáutica espacial ha incidido en que ocurra una generación de datos cada vez mayor. El capítulo 5 aborda los elementos relacionados con el análisis exploratorio de datos con alta densidad espacial. El problema se resuelve mediante el desarrollo, implementación y utilización de técnicas de visualización de datos multiparamétricos integradas dentro de sistemas de información geográfica, donde los grandes volúmenes de datos son manipulados con formatos de datos científicos integrados en sistemas de información geográfica.

La cuarta parte abordada en el capítulo 6, describe un conjunto de herramientas y algoritmos desarrollados sobre formatos de datos científicos para facilitar la utilización de las técnicas de visualización en las partes segunda y tercera.

Por todo lo anterior se formula el problema de investigación siguiente: No existen herra-

mientas computacionales para la visualización efectiva de grandes volúmenes de datos espacio-temporales, que mediante la aplicación de un modelo para facilitar y guiar el desarrollo de herramientas de geovisualización permita el análisis y la extracción de conocimiento, a través de la visualización de múltiples variables de series temporales simultáneamente.

El presente trabajo se orienta hacia el desarrollo de herramientas de Geovisualización, que permitan la extracción de conocimiento a partir de datos espacio-temporales con múltiples variables, tanto cuando se cuenta con datos con baja densidad espacial y amplios en el tiempo, como con datos con una alta densidad espacial amplios en el tiempo.

1.2. Objetivos

Para la solución de los problemas tratados anteriormente se trazó el siguiente **objetivo general**: Desarrollar un modelo que permita integrar herramientas de visualización en sistemas de información geográfica para el análisis de grandes volúmenes de datos espacio-temporales, brindando la posibilidad de analizar simultáneamente múltiples variables geo-referenciadas, que pueden representar series temporales.

Para la consecución de este objetivo general se plantean los siguientes **objetivos específicos** de la investigación:

1. Estudiar sistemáticamente las técnicas de visualización que se puedan modificar para su integración con sistemas de información geográfica. Implementar y adaptar estas técnicas de visualización para el análisis de grandes volúmenes de datos espacio-temporales.
2. Desarrollar un modelo que permita integrar técnicas de visualización en sistemas de información geográfica que facilite la manipulación e integración de datos geográficos y multiparamétricos.
3. Comprobar la viabilidad de la utilización del modelo para la extracción de conocimiento, reconocimiento de patrones, tendencias, anomalías y relaciones entre múltiples variables en diferentes áreas de aplicación, así como la formulación y corroboración de hipótesis. Utilizar el criterio de expertos para validar el modelo y las herramientas desarrolladas.

En la realización de este trabajo se aplicaron diferentes métodos de trabajo científico. Específicamente se empleó el método análisis-síntesis para procesar, integrar, interpretar y valorar la información consultada; el método sistémico para presentar el objeto de investigación en su integridad, reflejando sus componentes y nexos de funcionalidad existentes entre ellos; y el método de modelación y síntesis para la concepción de los sistemas desarrollados como contribución de este trabajo. La observación fue empleada para apreciar los resultados publicados en la literatura y determinar las deficiencias que existen en estos. El criterio de expertos y usuarios es utilizado para evaluar la validez de las herramientas desarrolladas en la solución de problemas reales.

1.3. Contribuciones

La producción científica asociada a esta tesis, está relacionada con las siguientes tareas de investigación realizadas:

- Estudio del uso de sistemas de información geográfica basados en software libre para la visualización de datos meteorológicos. Este trabajo permitió valorar las capacidades de los principales sistemas de información geográfica libres para la visualización de datos. El resultado fue presentado y publicado en un evento internacional en Cuba (Vázquez-Rodríguez *et al.*, 2009c).
- Estudio de la factibilidad de la aplicación de técnicas de visualización de datos multiparamétricos para el análisis visual de datos meteorológicos. Este trabajo permitió demostrar la validez de la idea de la utilización de técnicas de visualización de datos multiparamétricos para realizar análisis en la rama de la meteorología. El resultado fue presentado y publicado en un evento internacional en Cuba (Vázquez-Rodríguez *et al.*, 2009a) y en el seminario internacional de doctorado en Softcomputing (Vázquez-Rodríguez *et al.*, 2009d).
- Implementación de algunas de las técnicas de visualización de datos multiparamétricos. Esta tarea fue realizada y publicada en el Congreso Cubano de Reconocimiento de Patrones en el año 2009 (Vázquez-Rodríguez *et al.*, 2009b) y en un evento internacional de descubrimiento de conocimiento y aprendizaje automatizado desarrollado en la ciudad de Santa Clara por académicos de Cuba y Bélgica (Vázquez-Rodríguez *et al.*, 2010c).
- Creación de un módulo de minería visual de datos para el sistema de información geográfica gvSIG. Para la realización de esta tarea se juntaron todas las técnicas de visualización implementadas y se integraron en el sistemas de información geográfica gvSIG. Este resultado fue publicado en la revista *Anuario do Instituto de Geociências* (Vázquez-Rodríguez *et al.*, 2013b) y presentado en un evento internacional en Valencia, España (Vázquez-Rodríguez *et al.*, 2010d); ambas publicaciones se encuentran indexadas en prestigiosas bases de datos bibliográficas.
- Uso de técnicas de visualización de datos multiparamétricos para el análisis de datos espacio-temporales en sistemas de información geográfica de escritorio. Tarea que se publicó en una revista cubana (Vázquez-Rodríguez *et al.*, 2010b) y en varios eventos internacionales en Cuba (Vázquez-Rodríguez *et al.*, 2011a,b).
- Desarrollo de un modelo que permite integrar herramientas de visualización en sistemas de información geográfica para el análisis de grandes volúmenes de datos espacio-temporales, brindando la posibilidad de analizar simultáneamente múltiples variables espaciales que representan series temporales. Esta tarea constituye la principal novedad científica de este trabajo. Los resultados fueron publicados en la *Revista Técnica de la Facultad de Ingeniería de la Universidad del Zulia* (Vázquez-Rodríguez *et al.*, 2015); revista que

figura en el listado de revistas del *Journal Citation Report*.

- Extensión del módulo de visualización de datos multiparamétricos de gvSIG con nuevas técnicas. Los resultados de esta tarea fueron presentados y publicados en varios eventos internacionales en Cuba, como Informática y Geociencias 2013 (Vázquez-Rodríguez *et al.*, 2013a,c)
- Integración de formatos de datos científicos en sistemas de información geográfica. Esta tarea de investigación fue desarrollada y publicada en Cuba en el evento internacional Geociencias 2013 (Vázquez-Rodríguez *et al.*, 2013d).

A partir del planteamiento del problema y como resultado del análisis de la literatura consultada, se formuló la hipótesis siguiente:

La creación de un modelo que permita integrar herramientas de visualización en sistemas de información geográfica para el análisis de grandes volúmenes de datos espacio-temporales manipulados con formatos de datos científicos, que brinde la posibilidad de analizar simultáneamente múltiples variables espaciales, las cuales representan series temporales, facilitará el reconocimiento de patrones, tendencias, anomalías y correlaciones entre múltiples variables espacio-temporales.

Por todo lo anterior el trabajo que se propone tiene valor teórico, práctico y metodológico.

El valor teórico o la novedad científica del presente trabajo consiste en:

La creación de un nuevo modelo, que ha conducido al desarrollo de un nuevo método de exploración de secuencias de datos espaciales en sistemas de información geográfica (que pueden representar series temporales). Este resultado está avalado, entre otras cosas, por dos premios provinciales de la Academia de Ciencias de Cuba y varias publicaciones en revistas y eventos internacionales.

El valor práctico consiste en la obtención de un conjunto de herramientas de *software*, como resultado de la implementación del modelo propuesto. Se cuenta con un total de cinco registros informáticos en el Centro Nacional de Derecho de Autor de Cuba, que son listados en el anexo 3. Todas las soluciones propuestas están basadas en software libre y pueden ser generalizadas en otras áreas de aplicación.

Por su parte, el valor metodológico consiste en el desarrollo y aplicación sistémica de tratamiento y análisis de datos espacio-temporales mediante el uso de técnicas de visualización de datos multiparamétricos soportados sobre sistemas de información geográfica.

1.4. Estructura de la tesis

Este documento ha sido estructurado en siete capítulos:

En el primer capítulo se hace una introducción del contenido de la tesis, se establecen los objetivos y se presentan las principales tareas de investigación trazadas. En el segundo capítulo se

presentan los referentes teóricos que se consideran relevantes relacionados con la visualización científica, los sistemas de información geográfica y el análisis exploratorio de datos, así como la integración entre estas disciplinas.

En el tercer capítulo se realiza la propuesta conceptual del modelo. El capítulo 4 se concentra en el análisis exploratorio de datos con baja densidad espacial, mientras que el capítulo 5 se enfoca en el análisis exploratorio de datos con alta densidad espacial. El capítulo 6 se dedica a la descripción de un grupo de herramientas para el soporte de formatos de datos científicos en sistemas de información geográfica. En el capítulo 7 se hace una validación de los resultados y por último se arriban a conclusiones y se presentan temas abiertos para trabajos futuros.

2 Visualización científica y sistemas de información geográfica para el análisis exploratorio de datos

2.1. Visualización Científica

La ciencia ha desarrollado diversos métodos para la obtención de información, y uno de ellos se basa en la creación de imágenes a partir de los datos. Este método, conocido como visualización, ha sido utilizado como vía natural para mostrar información (Hansen y Johnson, 2005). La visualización por computadora es un proceso de distribución de las representaciones hechas por la computadora a representaciones perceptibles, mediante técnicas de codificación con el objetivo de maximizar el entendimiento y la comunicación con los seres humanos (Hansen y Johnson, 2005).

La visualización científica transforma los datos científicos y abstractos en imágenes. Esta forma especial de visualización pretende encontrar una representación visual apropiada para un conjunto de datos que permite mayor efectividad en el análisis y la evaluación de los mismos, y se destaca, además, por simplificar el análisis, la comprensión y la comunicación de modelos, conceptos y datos en la ciencia y la ingeniería (Hansen *et al.*, 2014).

La visualización de datos permite alcanzar diferentes metas. La naturaleza del objetivo que se desee está en relación directa con el conocimiento que se tenga sobre los datos inicialmente. Los objetivos pueden ser los siguientes (Abello y Korn, 2002):

- Análisis exploratorio.
- Análisis confirmativo.
- Presentación de información.

En el análisis exploratorio se tiene un conjunto de datos sin una hipótesis específica. Estos datos se someten a un proceso de búsqueda interactiva de información, que genera como resultado una visualización que soporta una hipótesis sobre el conjunto de datos. El reto es lograr

descubrir resultados completamente inesperados, sin conocimiento previo sobre los datos. En el análisis confirmativo se tiene un conjunto de datos sobre los que se plantea una hipótesis. Se realiza un procesamiento de dichos datos que genera una visualización mediante la cual se pueda validar o refutar la hipótesis que se tenía de ellos. La presentación de información parte de hechos que son fijos *a priori* y en los cuales se desea enfatizar y mostrar con extrema calidad (Keim *et al.*, 2006).

Los análisis exploratorio y confirmativo involucran también el análisis estadístico de los datos, la simulación y la educación. Dentro del área del análisis existen aplicaciones para el control de la calidad, el análisis de esfuerzos, las proyecciones financieras, entre otras; esta última podría considerarse dentro del área de simulaciones, junto con la de modelos atmosféricos. En cuanto a la educación, se tienen desde demostraciones matemáticas hasta modelos de física cuántica y planetarios (Keim *et al.*, 2006).

El análisis visual de datos es un nuevo enfoque que se puede utilizar con cualesquiera de los objetivos anteriores (Keim *et al.*, 2008a). Como se ha mencionado anteriormente, integra tanto la percepción humana como los métodos computacionales automáticos, lo que permite una mejor comprensión y análisis de grandes y complejos conjuntos de datos.

2.2. Técnicas de visualización científica

Para agrupar y clasificar las diferentes técnicas de visualización científica existentes se han empleado diversos enfoques. Un enfoque establecido para clasificar las técnicas es a través del tipo de dato sobre el que opera. Por tipo de dato se entiende al tipo que pertenecen los atributos o variables. Atendiendo a este criterio se encuentran las siguientes categorías (Theisel, 2000; Hansen y Johnson, 2005):

- Técnicas de visualización para datos volumétricos.
- Técnicas de visualización para fluidos.
- Técnicas de visualización para datos multiparamétricos.
- Técnicas de visualización de la información.

Para especificar los datos existen diversos enfoques, los cuales permiten definir una serie de características de los datos como son la dimensionalidad, la estructura y el nivel de medición.

Los datos volumétricos representan una malla de tres dimensiones donde cada punto tiene asociado un valor. En general los datos se definen como un conjunto S de muestras, donde cada elemento que pertenece a S es un vector de la forma (x, y, z, v), que contiene las coordenadas espaciales y un elemento que es un escalar (Theisel, 2000; Hansen y Johnson, 2005).

Los campos vectoriales representan una malla de dimensión menor o igual que tres, donde cada punto está relacionado con un vector. Una de las áreas de mayor uso de los campos vectoriales es para representar datos de fluidos (Hansen y Johnson, 2005).

Los datos multiparamétricos son aquellos en que el número de variables relacionadas con cada observación es mayor o igual que dos. Estas variables pueden ser cuantitativas o cualitativas y a su vez ordinales o nominales (Hansen y Johnson, 2005).

En algunas aplicaciones los datos presentan una estructura que no concuerda con ninguna de las anteriores o que sencillamente no puede ser definida con exactitud. A estos datos se les suele llamar información y entre las principales se identifican estructuras como árboles, grafos e hipertexto (Theisel, 2000; Keim, 2002b; Hansen y Johnson, 2005).

Keim (2002b) introdujo la minería visual de datos como una combinación de técnicas de minería de datos tradicionales con técnicas de visualización de información. La misma fusiona técnicas poderosas de análisis automático con las capacidades perceptivas y cognitivas de los humanos. Las técnicas de minería visual de datos son particularmente adecuadas cuando:

- se necesita explorar grandes cantidades de datos,
- hay poco conocimiento acerca de los datos o los objetivos de análisis son ambiguos,
- los datos son complejos o contienen ruido.

La minería visual de datos se puede ver como una exploración interactiva, en la que el usuario está directamente involucrado en el proceso de análisis, por lo tanto se hace necesaria la comprensión de los algoritmos básicos automáticos y los parámetros que estos necesitan para funcionar correctamente. De esta forma, la exploración se puede acelerar normalmente, se logran mejores resultados que los que se obtienen en el proceso de ejecución automático puro, y aumenta así la confianza en los resultados. El objetivo no es reemplazar métodos automáticos por visuales, sino acoplar al proceso de exploración toda una variedad de métodos que faciliten el análisis.

Esta investigación se centró en el desarrollo de técnicas de visualización de datos multiparamétricos. Es por ello que se hace mayor énfasis en este tipo de técnicas, las cuales son descritas a continuación.

2.3. Técnicas de visualización para datos multiparamétricos

Existe una serie de problemas en que cada punto de dato contiene más de un atributo, estos atributos pueden ser fechas, precios o valores descriptivos. A este tipo de datos se les llama multiparamétricos y se encuentran generalmente en aplicaciones de minería de datos, estadísticas e inteligencia artificial (Keim, 2002a). Los datos multiparamétricos, también llamados multidimensionales o datos n-dimensionales, consisten en un número de m registros donde cada uno está definido por un vector de v valores. Estos datos pueden ser vistos como una matriz de $m * v$, donde cada fila representa un registro y cada columna una observación, variable o dimensión (Ward, 2002).

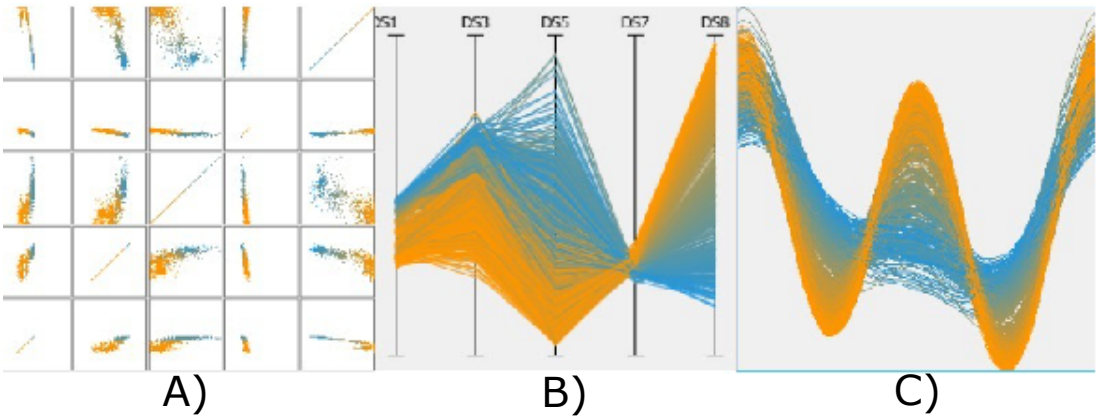

Figura 2.1 Ejemplos de técnicas de visualización geométricas. A) Matrices de diagramas de dispersión. B) Coordenadas paralelas. C) Gráfico de Andrews.

El objetivo fundamental de los métodos de visualización para datos multiparamétricos es lograr que las representaciones revelen correlaciones o patrones entre las variables (Eick, 2000; Keim, 2002b). Con este fin existe actualmente una amplia gama de técnicas de visualización, para las cuales se han creado además diversas mejoras. Las técnicas pueden ser clasificadas en geométricas, basadas en iconos, basadas en píxel y proyecciones (Keim, 2002a), entre otras.

2.3.1. Técnicas geométricas

Las técnicas geométricas se basan en el establecimiento de una relación entre los datos correspondientes a los atributos y un espacio geométrico (Mazza, 2009). Dicho de otra forma, son las técnicas que utilizan elementos como puntos, líneas o curvas como propiedades visuales para representar los datos (Theisel, 2000; Keim, 2002a). Existe un gran número de ellas, las visualizaciones geométricamente transformadas pretenden encontrar patrones *interesantes* en conjuntos de datos multidimensionales.

Esta clase de métodos incluye: las técnicas de exploración estadísticas -por ejemplo, obsérvese (Theus, 2005)-, tales como matrices de diagramas de dispersión (obsérvese la figura 2.1-A) (Andrews, 1972; Cleveland, 1993) y las técnicas que pueden incluirse bajo el término de búsqueda de la proyección o *projection pursuit* (Huber, 1985). Otras de las técnicas geométricas son *Prosections Views* (Furnas y Buja, 1994; Spence *et al.*, 1995), super rebanadas o *Hyper Slices* (Wijk y Liere, 1993), Parahistogramas (Ong y Lee, 1996), *Landscapes* (Wright, 1995), coordenadas de estrella (Kandogan, 2000), pero hay tres que sobresalen por su generalidad y gran uso, mencionadas anteriormente: matriz de diagramas de dispersión (obsérvese la figura 2.1-A), coordenadas paralelas (obsérvese 2.1-B) (Inselberg y Dimsdale, 1990; Cleveland, 1993; Keim, 2002b; Cui *et al.*, 2006) y gráfico de Andrews (obsérvese la figura 2.1-C) (Andrews, 1972).

Diagramas de dispersión

El diagrama de dispersión es una técnica sencilla muy utilizada. Su forma más simple se manifiesta cuando los datos poseen solo dos dimensiones, y entonces la técnica consiste en trazar dos ejes de coordenadas y utilizar los valores de las dimensiones como puntos (x, y) de R^2, de donde resulta un gráfico en el cual se encuentran dispersos los puntos de datos. Para visualizar datos de más de dos dimensiones pueden utilizarse proyecciones, las cuales provocan pérdida de información debido a la reducción de la dimensión (Theisel, 2000; Keim, 2002b; Hansen y Johnson, 2005).

Para datos multiparamétricos es muy frecuente utilizar matrices de diagramas de dispersión. Las matrices resultantes son cuadradas y el elemento (i, j) de la matriz es un diagrama de dispersión de la dimensión i y la j. El diseño evita la pérdida de información, pero en cambio los análisis complejos son engorrosos. Una deficiencia adicional es que la diagonal principal de la matriz es subutilizada. Algunos trabajos actuales están encaminados a aprovechar mejor esta región de la representación (Cui *et al.*, 2006).

Coordenadas paralelas

Las coordenadas paralelas son un método de visualización diseñado para crear una representación en 2D de datos multidimensionales sin pérdida de información. La técnica fue introducida por Inselberg y Dimsdale (1990). En ella se visualiza una tupla de datos $(x_1, x_2, ..., x_n)$ como una línea poligonal, que conecta los puntos $x_1, x_2, , x_n$ en n ejes y paralelos (obsérvese la figura 2.1-B). Para volúmenes de datos suficientemente grandes la visualización de la técnica puede llegar a ser confusa. Una posible solución consiste en una extensión en 3D, donde el plano xy representa la versión en 2D de las coordenadas paralelas, mientras la dimensión z representa la densidad de los eventos (Streit *et al.*, 2006).

Table lens

Table lens es una técnica de visualización inspirada en las aplicaciones de hojas de cálculo, pero a diferencia de estas los datos se representan mediante barras horizontales en lugar de valores numéricos (obsérvese la figura 2.2-A). Los datos se representan en una matriz donde los atributos se representan en columnas y cada instancia de los datos se reporta en una fila (Rao y Card, 1994; Mazza, 2009). Table Lens ha sido caracterizada como una herramienta efectiva para comprender las características de datos numéricos y categóricos en conjuntos de datos multivariados (Pirolli y Rao, 1996).

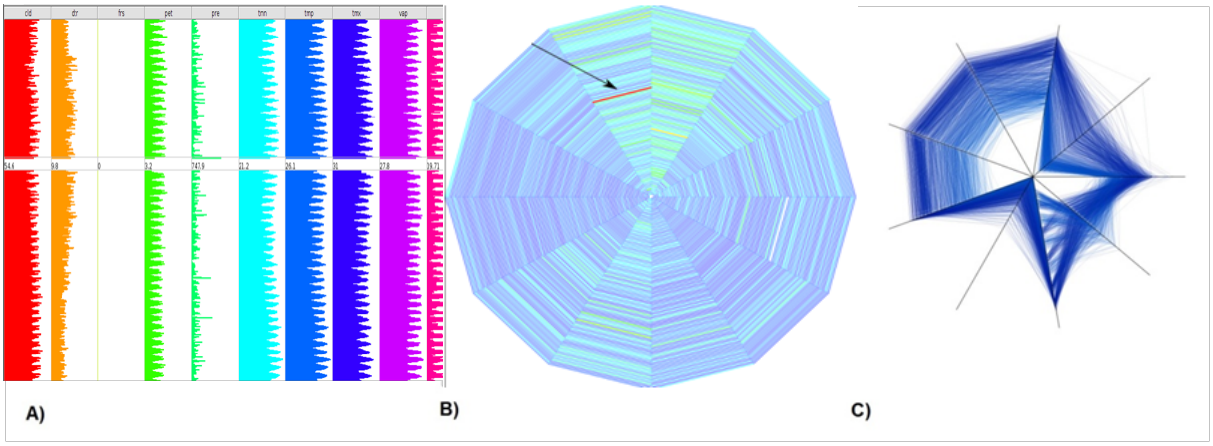

Figura 2.2 Otras técnicas geométricas. A) Table lens (Pirolli y Rao, 1996). B) Vista circular (Keim *et al.*, 2004). C) Coordenadas de estrella.

Vista circular

El proceso de visualización en una vista circular según Keim *et al.* (2004) consiste en la división de un círculo en un número de segmentos en dependencia del número de variables del conjunto de datos a mostrar. Cada segmento a su vez se divide en subsegmentos, que a menudo indican momentos de tiempo (obsérvese la figura 2.2-B).

Coordenadas de estrella

Las coordenadas de estrella consisten en la disposición de los ejes coordenados de forma circular en un plano bidimensional, con ángulos iguales entre los ejes y origen en el centro de la circunferencia (Kandogan, 2000). El sistema de coordenadas de estrella es un sistema coordenado curvilíneo, que puede ser convertido a coordenadas cartesianas mediante la definición de un punto bidimensional $O_n(x, y) = (o_x, o_y)$ que representa el origen y una secuencia de n vectores bidimensionales los ejes:

$$A_n = \langle a_1, a_2, \ldots, a_i, \ldots, a_n \rangle$$

La conversión de un elemento D_j de un conjunto de datos D hacia un punto bidimensional en las coordenadas cartesianas está determinada por la suma de todos los vectores unitarios $u_i = (u_{xi}, u_{yi})$ en cada coordenada, multiplicada por el valor del elemento de datos de esa coordenada:

$$P_j(x, y) = \left(ox + \sum_{i=1}^{n} u_{xi} \times (d_{ji} - \min_i), oy + \sum_{i=1}^{n} u_{yi} \times (d_{ji} - \min_i) \right)$$

donde:

$$D_j = (d_{j0}, d_{j1}, \ldots, d_{ji}, \ldots, d_{jn}),$$

$$\left|\frac{a_i}{\text{máx}_i - \text{mín}_i}\right|,$$

$$\min_i = \text{mín}\left\{d_{ji}, 0 \leq j < \left|D\right|\right\}, \max_i = \text{máx}\left\{d_{ji}, 0 \leq j < \left|D\right|\right\}.$$

En la figura 2.2-C) se muestra un gráfico basado en coordenadas de estrella con datos de 9 variables. En este caso particular se ha definido un gradiente de color, de acuerdo con los valores de una de las variables representadas.

2.3.2. Técnicas basadas en iconos

Las técnicas basadas en iconos visualizan datos multidimensionales mediante la asignación de cada objeto de datos sobre valores de los parámetros en pequeñas gráficas primitivas. Normalmente los valores de los atributos están representados por la x e y posición del icono, así como la longitud, el ángulo o la forma de algún componente cónico. Para lograr un buen resultado, los componentes dentro de un icono deben ser distinguibles, iconos separados deben ser claramente identificables, y los iconos deben ser percibidos como distintos si difieren en algunos de los componentes.

Las técnicas basadas en iconos tienen dos parámetros que las caracterizan. El primero es el tipo de figura que representará cada observación, o sea, la forma del icono; el segundo parámetro es la forma en que se definirá la posición de cada icono en la imagen (Theisel, 2000; Ward, 2002).

Estas técnicas no sufren de pérdida de información. La pérdida de información se logra evitar al realizar una proyección de las dimensiones a los diferentes rasgos del icono (Theisel, 2000).

Las técnicas basadas en iconos son recomendadas cuando el número de dimensiones oscila entre diez y quince, y el número de mediciones de las mismas es alto. Estas técnicas se pueden utilizar con una referencia espacial.

Ejemplos de técnicas basadas en icono son: caras de Chernoff (obsérvese la figura 2.3-A)(Chernoff, 1973), los iconos de flechas o *Needle Icons* (Keim, 2000; Abello y Korn, 2002), figuras con palillos o *Stick Figure Icons* (obsérvese la figura 2.5)(Pickett, 1970; Pickett y Grinstein, 1988), iconos de colores (Levkowitz, 1991; Keim y Kriegel, 1994), iconos de barras *Tile Bars* (Hearst, 1995), codificación de formas y colores *shape coding* (obsérvese la figura 2.3-D)(Beddow, 1990), *profile glyphs* (obsérvese la figura 2.3-C) (Chen *et al.*, 2008) y campo de estrellas (obsérvese la figura 2.3-B) (Eick, 2000; Keim, 2002b; Ward, 2002; Xie *et al.*, 2006).

Caras de Chernoff

Este tipo de diagrama surge como respuesta a la facilidad que tienen las personas para reconocer y clasificar a otras personas por sus rostros, gracias a la capacidad de percepción (Chernoff,

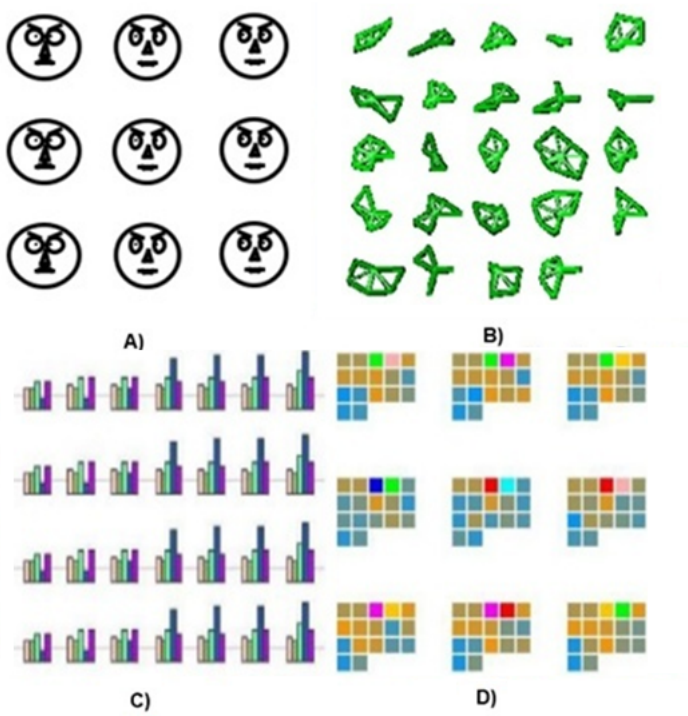

Figura 2.3 Ejemplos de técnicas de visualización basadas en iconos. A) Caras de Chernoff. B) Campo de estrellas. C) Iconos de barras. D) *Shape coding*.

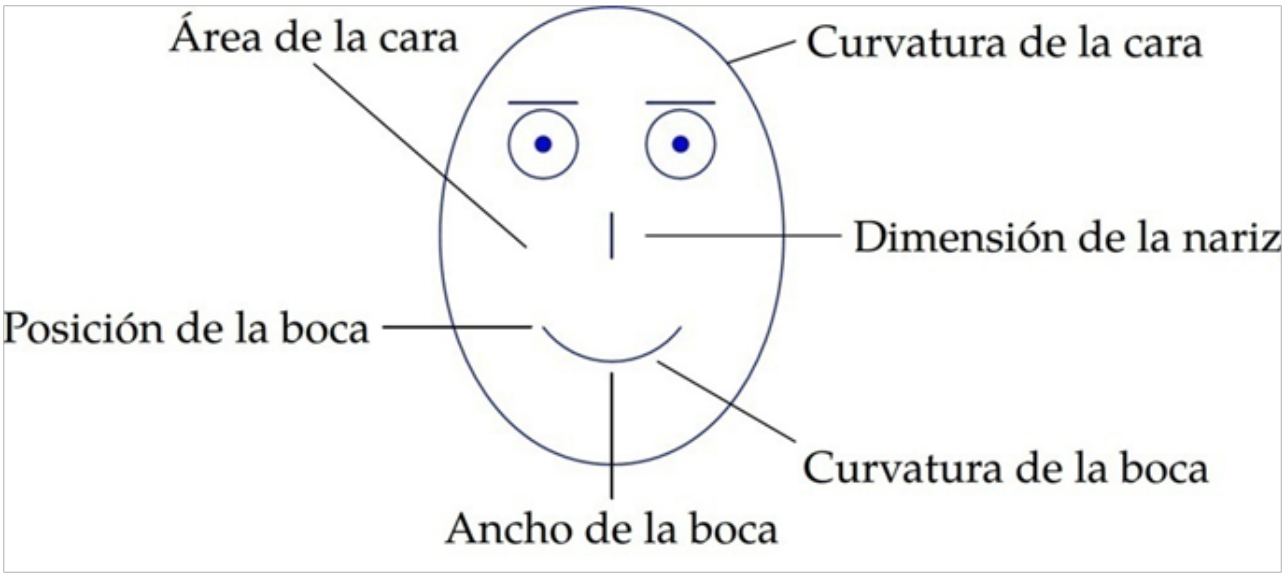

Figura 2.4 Algunos parámetros configurables en una cara de Chernoff

1973). Esta capacidad de percepción es la que se desea explotar con los diagramas de Chernoff, los cuales usan representaciones de rostros de tipo trazos, y mediante el tamaño de los ojos, nariz, cejas y boca, agregándosele su forma o curvatura, incluso de la misma cabeza, permiten combinar los diferentes atributos de los datos multivariados en un único símbolo. (Obsérvese algunos parámetros configurables en la figura 2.4)

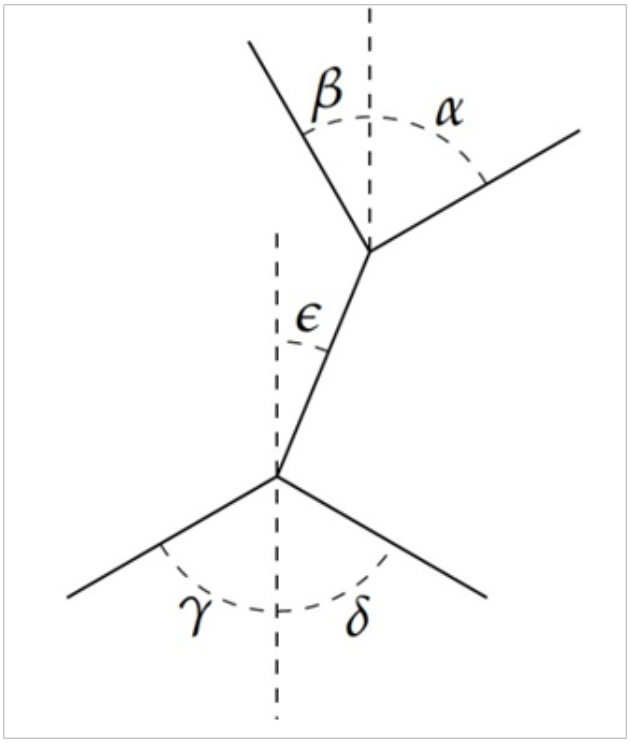

Figura 2.5 Parámetros modificables en figuras con palillos.

Figuras con palillos

La visualización de figura con palillos fue descrita por primera vez por Pickett y Grinstein (1988). Es rica en conceptos y aplicaciones prácticas (Wong y Bergeron, 1994). En su forma más básica se representa cada dato multivariado del conjunto por un icono gráfico, cuyas características visibles son controladas por los valores de cada una de las variables.

El icono original es una figura formada por cinco palillos (cuatro miembros y un cuerpo) con ángulos controlables en las extremidades (obsérvese la figura 2.5). De esta forma podrían representarse hasta 5 variables; no obstante existen variantes que permiten modificar, además, el tamaño, el grosor o el color de los palillos.

2.3.3. Técnicas orientadas a píxel

La visualización de un conjunto de datos de gran tamaño resulta un reto para técnicas geométricas y basadas en iconos. Al graficarlos suele surgir desorden en la imagen, originado por el tamaño de la figura que representa una observación simple. A partir de esta idea resulta lógico concluir que al minimizar el espacio que ocupa un solo punto de dato en la imagen se mejoraría la percepción visual (Andrews, 2005; Hansen y Johnson, 2005).

Para lograr la maximización del número de elementos a representar, algunas técnicas utilizan los píxeles de la pantalla como unidades básicas de representación (Mazza, 2009). El procedimiento consiste en relacionar cada valor de una dimensión a un color y agrupar los píxeles

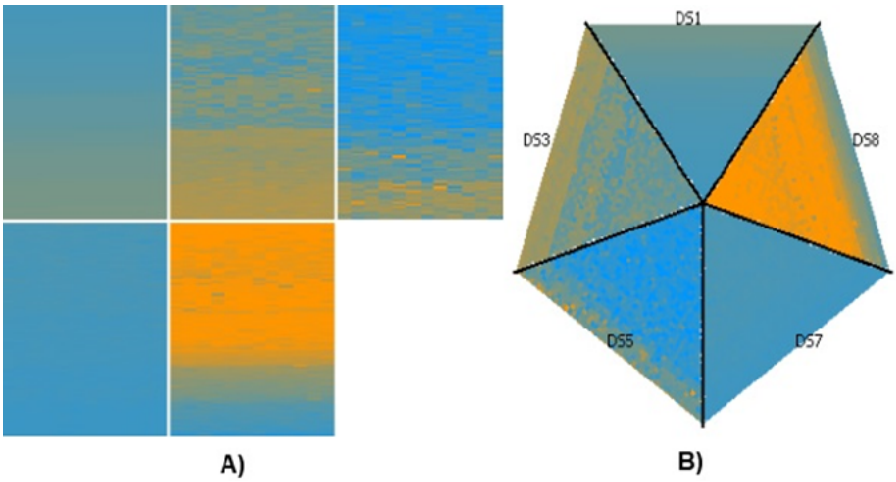

Figura 2.6 Ejemplos de técnicas de visualización orientadas a píxel. A) Patrones recursivos. B) Segmentos de círculo.

de cada dimensión en áreas adyacentes (Keim, 2002a). De esta manera una computadora con una resolución de pantalla de $1024 * 800$ píxeles podría potencialmente representar 819200 elementos de un conjunto de datos univariado.

Este tipo de técnicas utiliza diferentes modos de posicionamiento de los píxeles para lograr diferentes objetivos. Colocar los píxeles en la forma apropiada ofrece la posibilidad de observar información sobre correlaciones, dependencias y regiones trascendentales. Dos de los modos de posicionamiento de los píxeles son los patrones recursivos (obsérvese la figura 2.6-A) (Keim *et al.*, 1995) y los segmentos de círculo (obsérvese la figura 2.6-B) (Ankerst *et al.*, 1996; Keim, 2000). Otros ejemplos de técnicas orientadas a píxel son la de espiral (Keim y Kriegel, 1994), y de ejes (Keim y Kriegel, 1994).

Patrones recursivos

La técnica orientada a píxel patrones recursivos tiene una manera especial de interactuar con los datos, que permite la definición de diferentes niveles de recursividad. Está particularmente dirigida a representar un conjunto de datos con un orden natural de acuerdo con un atributo; propiedad que la convierte en una opción para problemas de series de tiempo. Una posibilidad simple que provee la técnica es la de organizar los puntos de datos de izquierda a derecha, línea por línea o columna por columna. Una vía posible de mejorar la visualización es la organización de los píxeles en pequeños grupos y organizar los grupos para formar un patrón global.

Esta estrategia corresponde a un planteamiento en dos fases con un patrón de primer orden formado por la agrupación de los píxeles y un patrón de segundo orden formado por el orden global. Al tomar los resultados de la estructura de segundo orden como el elemento básico de construcción de una estructura de tercer nivel, puede realizarse la introducción de un tercer patrón. Este proceso puede ser repetido hasta un nivel arbitrario formando un esquema general recursivo, el cual puede ser distribuido de dos formas para cada nivel de recursividad: línea a

```
DRAW(x, y, level)
 1  if x = 0
 2     then SET-PIXEL(x, y, color)
 3     else  for h ← 1 to height[level]
 4                 do if h mód 2 = 0
 5                       then for w ← 1 to width[level]
 6                                  do DRAW(x, y, level −1)
 7                                     x ← x + ∏_{i=1}^{level-1} w_i
 8                       else  for w ← 1 to width[level]
 9                                  do x ← x − ∏_{i=1}^{level-1} w_i
10                                     DRAW(x, y, level − 1)
11                    y ← ∏_{i=1}^{level-1} h_i
```

Figura 2.7 Algoritmo de patrones recursivos

línea (*line by line*), donde las estructuras de cada orden se posicionan en la imagen de izquierda a derecha, o intercalando el sentido (*back and fort*), donde las estructuras de cada orden se posicionan intercalando el sentido; es decir para una línea se posicionan de izquierda a derecha y en la siguiente línea de derecha a izquierda (Keim *et al.*, 1995).

La misma secuencia básica se hace en todos los niveles de recursión con la única diferencia de que los elementos básicos situados en el nivel i son los patrones resultantes de las ubicaciones del nivel $i-1$. Si w_i es el número de elementos ubicados de izquierda a derecha en el nivel i y h_i es el número de filas en el nivel i, entonces el patrón en el nivel i consiste de $w_i * h_i$ nivel$(i-1)$-patrones, y el máximo número de píxeles que pueden ser representados en el nivel k está dado por la productoria desde $i = 1$ hasta k de $w_i * h_i$ (Keim *et al.*, 1995).

El algoritmo $Draw$ (obsérvese el listado de la figura 2.7) permite realizar la visualización de patrones recursivos. Inicialmente el algoritmo se ejecuta con una llamada a la función $Draw(0, 0, \text{MAX-LEVEL})$ con el ancho y alto de todos los niveles de recursividad almacenados en un arreglo previamente definido. La condición de parada es la llegada al nivel de recursión 0. Para los niveles $i(i \geq 1)$, el algoritmo dibuja w_i nivel$(i-1)$-patrones h_i veces; en dependencia de la opción *line by line* o *back and fort*, lo hace alternando o no el sentido.

Segmentos de círculo

Como su nombre lo indica, la idea fundamental de esta técnica es mostrar las dimensiones de los datos como segmentos de un círculo. Si el conjunto de datos consiste en n variables, el círculo es consecuentemente particionado en n segmentos. Los elementos de los datos dentro de cada segmento son organizados de un lado hacia el otro a través de la llamada línea de dibujo o $drawline$, ortogonal a la línea que divide las dos líneas del borde, los segmentos (obsérvese la figura 2.6-C). Cada vez que la línea de dibujo toca uno de las líneas del borde, la primera

```
FILL-SEGMENT(l-1, l-2)
 1  x, y, direction ← 1
 2  record-count ← INITIAL-PIXELS(l-1, l-2, x, y)
 3  while record-count < RECORD-ALL
 4      do while POINT-BETW-LINES(l-1, l-2, x, y) = TRUE ∧ record-count < RECORD-ALL
 5             do record-count ← record-count +1
 6                SET-PIXEL(x, y, color)
 7                COMPUTE-NEXT-POINT(draw-line, x, y, direction)
 8          MOVE(draw-line)
 9          COMPUTE-NEXT-POINT(draw-line, x, y, direction)
10          direction ← − direction
11          while POINT-BETW-LINES(l-1, l-2, x, y = FALSE
12              do COMPUTE-NEXT-POINT(draw-line, x, y, direction)
```

Figura 2.8 Algoritmo de segmentos de círculos.

se mueve en paralelo junto a la línea divisoria del segmento hacia el exterior del círculo y la dirección de la línea de dibujo cambia. Este proceso se repite luego para cada una de las variables restantes (Ankerst *et al.*, 1996; Keim, 2000; Hansen y Johnson, 2005).

El algoritmo $Fill - Segment$ (obsérvese el listado de la figura 2.8) se llama con las dos líneas del borde del segmento y se introduce en la subrutina $Initial - Pixels$. La función $Initial - Pixels$ dibuja los primeros píxeles de un segmento hasta que la siguiente línea de dibujo tenga al menos un píxel entre las dos líneas del borde. El valor de retorno de $Initial - Pixels$ es el número de píxeles dibujados hasta ahora. La función $Compute - Next - Point$ se mueve hacia delante por la línea de dibujo. La función $Point - Betw - Lines$ comprueba si un punto está aún en el segmento. Si el punto no está en el segmento, la línea de dibujo se mueve un píxel hacia el exterior en paralelo con la línea divisoria del segmento. La nueva línea de dibujo traza el píxel en la dirección opuesta a la anterior.

2.3.4. Otros tipos de técnicas. Representación del tiempo, técnicas basadas en ejes, SOM, etc.

Los datos que implican cambios a través del tiempo constituyen un reto para la visualización y el análisis de datos. Existen varias técnicas para la visualización de conjuntos de datos con componentes temporales. Sin embargo, dado que es difícil considerar todos los aspectos concernientes a la dimensión tiempo en una sola visualización, la mayoría de los métodos disponibles tratan únicamente casos específicos, principalmente la visualización de datos con un eje de tiempo lineal (Keim *et al.*, 2010).

No obstante, las series temporales a menudo muestran estructuras periódicas que no pueden ser apreciadas a plenitud con este tipo de visualizaciones. La espiral de tiempo provee un ejemplo

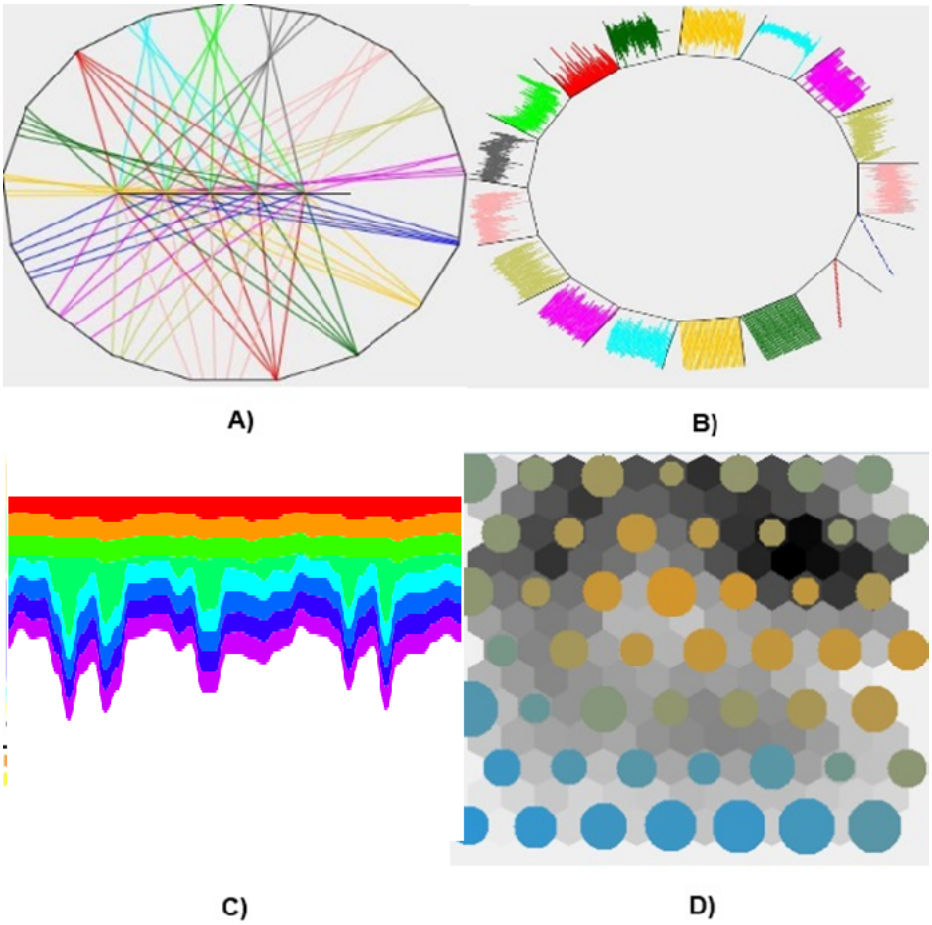

Figura 2.9 Ejemplos de otras técnicas de visualización. A) Rueda de tiempo. B) Combo temporal. C) Río temático. D) Mapas Auto-Organizados.

en que este tipo de patrón solo se hace evidente si se utiliza una visualización basada en una espiral de Arquímedes (Weber *et al.*, 2001).

Existen muchas otras formas de visualizar datos que permiten extraer información relevante relacionada con el tiempo. Ejemplo de estas son las técnicas basadas en ejes como la rueda de tiempo Time Wheel (obsérvese la figura 2.9-A) (Tominski *et al.*, 2004; Aigner *et al.*, 2008), su idea básica es presentar el eje de referencia (tiempo en este caso) en el centro de la pantalla, y circularmente organizar los ejes a su alrededor. Parcelas múltiples MultiComb (obsérvese la figura 2.9-B) (Abello y Korn, 2002): su objetivo básico es alinear los datos por parcelas de manera circular. Las parcelas de datos están formadas por la información de una variable con respecto al tiempo. Otros ejemplos son la visualización de río temático (Havre *et al.*, 2000) (obsérvese la figura 2.9-C) que es de gran utilidad para ver la variación de datos en una gran colección de información. Los cambios son mostrados en el contexto de una línea de tiempo. Los cambios en la imagen permiten al usuario discernir patrones más fácilmente y analizar la relación entre los datos.

Mapas Auto-Organizados

Los mapas auto-organizados (obsérvese la figura 2.9-D) o redes de Kohonen (SOM por sus siglas en inglés, *Self-OrganizingMap*) fueron introducidos por el profesor finlandés Teuvo Kohonen en los artículos (Kohonen, 1982, 1990; Kaski *et al.*, 1998). Un mapa auto-organizado es una herramienta que analiza datos en muchas dimensiones con relaciones complejas entre ellos y los presenta en una visualización sencilla en sólo dos dimensiones.

La propiedad más importante de SOM es que preserva las propiedades topológicas de los datos, es decir, que datos parecidos aparecen cercanos en la visualización. Este tipo de red posee un aprendizaje no supervisado competitivo. La red auto-organizada debe descubrir patrones regulares, elementos comunes, interrelaciones o categorías en los datos de entrada y adicionarlos a su estructura interna de conexiones (Brio y Molina, 2001). Brio y Molina (2001) dice que "en este caso que las neuronas deben autoorganizarse en función de estímulos (señales o datos) procedentes del exterior".

En el aprendizaje competitivo las neuronas compiten para efectuar una tarea determinada. La idea es que cuando se presente un patrón de entrada en la red neuronal, se active sólo una de las neuronas de la capa de salida. Esta es la neurona ganadora y las otras son perdedoras ante este estímulo o patrón.

Una red SOM está integrada por dos capas de neuronas. La capa de entrada y la capa de salida. La función de la capa de entrada es recibir y transmitir a la capa de salida la información externa. En la capa de salida la información procedente de la capa de entrada es procesada y es aquí donde se forma el mapa de rasgos. En la capa de entrada se tienen tantas neuronas como variables de entrada, mientras que en la capa de salida las neuronas se organizan como un mapa de dos dimensiones (Brio y Molina, 2001). En la figura 2.9-D se presenta el resultado de un agrupamiento realizado por SOM. Los círculos coloreados representan los grupos y el tamaño del círculo representa la cantidad de elementos que tiene cada clúster. Tamaño y colores similares de clústeres cercanos en el mapa, representan grupos similares. La coloración negra y blanca del fondo representa que clústeres separados por un fondo blanco son más parecidos a sus vecinos que los que están separados por un fondo oscuro.

La técnica SOM ha sido utilizada en múltiples estudios relacionados con el análisis visual de datos geográficos (Burns y Skupin, 2013; Sagl y Delmelle, 2014).

Las técnicas de visualización estudiadas en estos epígrafes son generales y se pueden aplicar en múltiples áreas. Además se pueden utilizar para analizar datos que tengan referencia espacial. Para esto es necesario integrarlas en sistemas de información geográfica. En el siguiente epígrafe se estudian algunos de los principales sistemas de información geográfica.

2.4. Sistemas de información geográfica

Aproximadamente un setenta por ciento de la información que se maneja en cualquier tipo de disciplina está georreferenciada. Es decir, que se trata de información que puede asignársele una posición geográfica, y por tanto información que viene acompañada de otra información adicional relativa a su localización. Esto demuestra que la situación es muy favorable para desarrollar aplicaciones que hagan uso de toda esta información. En una sociedad donde la información y la tecnología son dos de los pilares fundamentales, los sistemas de información geográfica constituyen la tecnología más relevante para el manejo de información geográfica, así como los elementos básicos que conlleva la gestión de todo aquello que presente una componente geográfica que pueda ser aprovechada.

Existen muchas definiciones de sistemas de información geográfica. En Bolstad (2005) se define un sistema de información geográfica como "un sistema computacional que ayuda en la recolección, mantenimiento, almacenamiento, análisis, visualización y distribución de información y datos espaciales". Otra definición es la de Jacobson *et al.* (2000), para quien un sistema de información geográfica es un elemento que permite "analizar, representar e interpretar hechos relativos a la superficie terrestre". El mismo autor, sin embargo, argumenta que "esta es una definición muy amplia, y habitualmente se emplea una más concreta. En otras palabras, un sistema de información geográfica es un conjunto de *software* y *hardware* diseñado específicamente para la adquisición, mantenimiento y uso de datos cartográficos".

Según Chorley (1987), es un sistema para capturar, almacenar, comprobar, manipular y visualizar datos que estén espacialmente referenciados a la tierra. A juicio de Clarke (1990) es un sistema automatizado para la captura, almacenamiento, composición, análisis y visualización de datos espaciales. Otra definición es en la que se considera como "un sistema de *hardware*, *software* y procesamiento diseñado para la captura, gestión, manipulación, análisis, modelado y visualización de datos espacialmente referenciados para resolver problemas complejos de planeamiento y gestión" (Cowen, 1989).

De manera similar, Star y Estes (1990) definen un sistema de información geográfica como un "sistema de información diseñado para trabajar con datos referenciados mediante coordenadas espaciales o geográficas. En otras palabras, un sistema de información geográfica es tanto un sistema de base de datos con capacidades específicas para datos georreferenciados, como un conjunto de operaciones para trabajar con esos datos.

Todas estas definiciones recogen el concepto fundamental de los sistemas de información geográfica en el momento en que fueron escritas, pero hoy día se hace necesario recoger otras ideas. La definición actual de un sistema de información geográfica debe fundamentarse sobre todo en el concepto de sistema, como elemento integrador que engloba un conjunto de componentes interrelacionados. Por todo lo antes expuesto, en esta investigación se asume la definición siguiente: un sistema de información geográfica es un sistema conformado por tecnología informática, personas e información geográfica, que está especialmente diseñado

para la captura, análisis, almacenamiento, edición y representación de datos georreferenciados (Korte, 2001; Olaya, 2011b).

El acelerado desarrollo de los sistemas de información geográfica, provocado en parte por la revolución de las nuevas tecnologías, los ha convertido en una herramienta de trabajo esencial para el análisis y resolución de diversos problemas que se presentan en empresas, industrias e instituciones sociales y gubernamentales. Su versatilidad ha permitido que puedan ser empleados en casi todas las actividades que poseen una componente espacial, convirtiéndose en una herramienta esencial en muchas áreas relacionadas con la gestión estratégica, incluidos también los procesos de análisis demográfico, protección del medio ambiente, y aplicaciones de urbanismo y gestión de recursos. En general, encontramos aplicaciones de los sistemas de información geográfica en tareas como:

- Localización: Acotar los límites y conocer las características de un lugar en el espacio.
- Condición: El cumplimiento o no de una de las condiciones impuestas al sistema.
- Tendencia: Comparar situaciones temporales o espaciales con características diferentes.
- Rutas: Hallar las rutas óptimas entre múltiples puntos.
- Modelos: Generación de modelos producto de fenómenos o simulaciones.

Los sistemas de información geográfica analizados en este epígrafe son ArcGIS, GRASS GIS, Quantum GIS, gvSIG, Open JUMP, uDIG y la biblioteca de algoritmos Sextante.

2.4.1. ArcGIS

El conjunto de productos de software en el campo de los sistemas de información geográfica conocido como ArcGIS, ha sido desarrollado y comercializado por la compañía ESRI. Este conjunto de productos contiene aplicaciones básicas propias de los sistemas de información geográfica para capturar, editar, analizar, manipular, diseñar, publicar y preparar para impresión mapas geográficos. Estas aplicaciones se enmarcan en ramas temáticas como ArcGIS Server, que sirve para publicar y gestionar mapas en la web, o ArcGIS Móvil que permite capturar y gestionar información en campo. Es un *software* propietario y para su utilización es necesario pagar una licencia (Shekhar y Xiong, 2008).

Una de las principales potencialidades de este sistema de información geográfica es la visualización que permite mostrar información geográfica, tal como: lugares, posiciones en el terreno, áreas urbanas y rurales, regiones y cualquier tipo de ubicaciones en terrenos determinados. Dicha información manipulada de manera integral en equipos de cómputo, marca la diferencia relacionada con el trabajo con información en planos y mapas impresos. De esta forma se pueden explorar, observar y analizar los datos teniendo en cuenta parámetros, relaciones y tendencias que presenta la información, como resultado se obtienen nuevas capas de información, mapas y bases de datos.

En ArcGIS se puede desplegar fácilmente gran cantidad de información geográfica, debido principalmente a lo intuitiva y amigable que es su interfaz gráfica. Cuando se cuenta con algunos conocimientos generales de sistemas de información geográfica, el aprendizaje con ArcGIS se agiliza, puesto que contiene una excelente ayuda en línea. Actualmente el ArcGIS Desktop incluye herramientas para manipular catálogos de metadatos (ArcCatalog), para construir mapas dinámicos e inteligentes que permiten visualizar patrones, tendencias y peculiaridades en sus datos (ArcMap), para la construcción de modelos 3D y animaciones (ArcGlobe y ArcScene). Incluye, además, herramientas para ejecutar cientos de algoritmos (ArcToolbox) que se pueden interrelacionar creando complejos flujos de datos para resolver problemas de gran envergadura (ModelBuilder) (Shekhar y Xiong, 2008).

2.4.2. GRASS GIS

GRASS (*Geographic Resources Analysis Support System*) es un sistema de información geográfica de propósito general de código abierto con una estructura en constante perfeccionamiento con el fin de adaptarse a las nuevas necesidades. Fue inicialmente concebido y desarrollado en 1982 por el laboratorio de investigación del cuerpo de ingenieros del ejército de los Estados Unidos (USA-CERL) para la gestión del territorio y la gestión medioambiental (GRASS, 2008). GRASS comenzó a difundirse en ámbitos educativos e instituciones públicas donde se desarrollaron numerosas aplicaciones alrededor de dicho sistema, hasta que en 1999 pasó a tener licencia del tipo GNU GPL. Está escrito en forma modular completamente por lo que se minimiza la sobrecarga, esto permite que los usuarios puedan ejecutar el sistema, o parte de este, en pequeños dispositivos portátiles con limitada RAM (Neteler *et al.*, 2012). Diversos estudios han demostrado que GRASS es una poderosa herramienta en muchas áreas de estudios y para resolver determinadas tareas dentro del ámbito científico.

GRASS dispone de un gran número de herramientas y utilidades. Originalmente estuvo más orientado al aspecto matricial (*raster*) de la información, aunque contaba con un potente editor de topología vectorial. Sin embargo, en las últimas versiones se ha potenciado el aspecto vectorial, y sobre todo la conexión externa a bases de datos. También ha experimentado una gran evolución en su interfaz de usuario, teniendo en cuenta que en las primeras versiones todo el control se hacía por medio de comandos tipo UNIX. Para la reciente versión 6.4 se incluye una nueva y moderna Interfaz Gráfica de Usuario la cual permite definir proyectos con bases de datos y toda su configuración a través de asistentes, construir consultas SQL, editar atributos, vistas en 3D, así como herramientas de georreferenciación (Neteler *et al.*, 2012). Esta nueva Interfaz Gráfica de Usuario fue escrita en Python. Otro gran avance ha sido la herramienta de visualización 3D (NVIZ), que se destaca por su potencia gráfica y las opciones de generación de salidas gráficas que permite. En la actualidad GRASS presenta más de 300 comandos que le dan una gran funcionalidad (Esparza Gil, 2014); además puede enlazarse directamente a varios *software* incluidos Quantum GIS y Sextante (una extensión de análisis para gvSIG), MATLAB

y otros.

GRASS está escrito en el lenguaje C con algunas funciones implementadas en los lenguajes $C++$ y Python. Es portable, dado que puede ser ejecutado en varios Sistemas Operativos (GNU/Linux, MacOSX, y MS-Windows son los sistemas soportados oficialmente), además puede ser descargado libremente desde Internet. Debido a que fue inicialmente diseñado para sistemas UNIX, tiene gran difusión en centros universitarios y de investigación. El proyecto GRASS representa un buen ejemplo de un modelo de desarrollo colaborativo entre comunidades de usuarios.

2.4.3. Quantum GIS

Quantum GIS (o QGIS) es un sistema de información geográfica tipo escritorio, muy intuitivo y fácil de utilizar que pretende ofrecer a usuarios con necesidades básicas un entorno sencillo y agradable. Su licencia es GNU, y por tanto se trata de código libre (Gray, 2008). Es multiplataforma y se pueden encontrar versiones para diferentes sistemas operativos: GNU/Linux, Unix, Mac OS y Microsoft Windows. Salió oficialmente como producto de la fundación OSGeo en 2008. Permite manipular formatos *raster* y vectoriales a través de las bibliotecas GDAL y OGR, así como bases de datos. Hasta no hace mucho, era uno de los pocos editores de PostGIS para la plataforma Windows y se destaca por su sencillez y velocidad.

Una de sus mayores ventajas es la posibilidad de usar Quantum GIS como GUI del sistema de información geográfica GRASS. Con la integración de estos dos software, se pueden explotar las capacidades de ambos para la visualización y el acceso a datos, así como también para el análisis *raster* y vectorial. Se utilizan como base las propias capacidades de GRASS, pero en un entorno de trabajo más amigable (Olaya, 2011c). QGIS está desarrollado en C++, usando la biblioteca Qt para su interfaz gráfica de usuario, actualmente permite la incorporación de nuevos módulos y funcionalidades implementadas en C++ y Python (Neteler, 2010).

2.4.4. gvSIG

gvSIG (Generalitat Valenciana) surge como un proyecto amparado por la Generalitat Valenciana de España que, a finales de 2003, promocionó un concurso para el desarrollo de un sistema de información geográfica con una serie de características propias que incluía entre otras, que fuera operable en múltiples plataformas, de código abierto, que permitiera extenderse mediante módulos, y fuera interoperable con formatos de otros programas (Autocad, Microstation, Arcview). Otro aspecto importante es que estuviera basado en estándares de la OGC (*Open Geospatial Consortium*) (Anguix y Carrión, 2005). El resultado ha sido una aplicación que ya tiene disponibles varias versiones al público y gran parte de las funcionalidades propias de los sistemas de información geográfica cubiertas, aunque se desarrolla constantemente. Las funciones básicas

que cualquier usuario desearía como diseño de impresión o soporte de formatos de imagen típicos están incorporadas sin necesidad de ningún módulo adicional.

gvSIG posee una jerarquía de clases bien estructurada para la incorporación de nuevas funcionalidades. Permite la lectura de varios formatos de datos geográficos y no geográficos en forma de tablas, así como la conexión con varias bases de datos. Este sistema de información geográfica posee las aplicaciones traducidas a veinte idiomas; toda la documentación está disponible en 5 idiomas, incluyendo español e inglés, por lo que se ha convertido en un sistema de información geográfica muy popular en el mundo hispano (Anguix, 2009). Se ha reportado su utilización en varios países europeos como Francia, Italia, Suiza, Austria, Reino Unido y Alemania, donde se encuentra la mayor comunidad de usuarios de gvSIG no hispanohablantes. Varias instituciones y universidades prestigiosas han utilizado esta aplicación, tal es el caso de la Agencia Espacial Europea y Oxford Archaeology. Varios países africanos también han realizado trabajos con gvSIG, pero su mayor uso se ha reportado en Iberoamérica.

Entre las funcionalidades que encontramos en gvSIG están:

- Acceso a formatos vectoriales, ráster, servicios remotos, bases de datos y tablas.
- Consultas.
- Geoprocesos.
- Representación vectorial y ráster.
- Redes.

2.4.5. Open JUMP

JUMP fue uno de los primeros sistemas de información geográfica gratuitos y por lo tanto ha servido de base a otros sistemas de información geográfica desarrollados, tanto libres como propietarios. Su origen está en Canadá, ya que nace como un proyecto patrocinado por una serie de instituciones canadienses (Steiniger y Hunter, 2012).

JUMP es un SIG modular escrito en Java y que basa su funcionalidad en módulos (plugins). De esta forma si queremos cargar cualquier tipo de imagen o dato vectorial sólo tenemos que encontrar o programar el módulo necesario. Lo mismo ocurre con cualquier funcionalidad adicional que se desee implementar: consultas, ediciones avanzadas, etc.

La interfaz de usuario es similar a la que proporciona ArcView, con una tabla de contenidos a la izquierda y una ventana central para el mapa. Es posible conectarse a servidores de cartografía WMS y existen *plugins* para numerosos formatos tanto de archivo como de servidores. Uno de los aspectos más interesantes son las herramientas de edición de que dispone para modificar datos vectoriales, así como herramientas básicas de geoprocesamiento (como zonas de influencia, intersecciones, uniones, etc). Existe también una versión muy prometedora para la edición y corrección de topología (*Jump Conflation Suite*) que se aproxima a funcionalidades de ArcMap en su versión de ArcINFO.

Actualmente han aparecido versiones internacionalizadas y varias páginas que albergan proyectos relativos a Jump, tanto para la creación de nuevas extensiones, como proyectos que basados en Jump procuran generar nuevos programas con funcionalidades más específicas.

Algunos de sus puntos fuertes radican en:

- Interfaz de usuario muy intuitiva.
- Soporte para una gran cantidad de formatos de datos, incluyendo conexiones a servidores de mapas y bases de datos.
- Buen punto de partida para la creación de proyectos personalizados debido a la documentación existente y a la facilidad de implementación de nuevas funciones.

Presenta algunas dificultades, como por ejemplo:

- La falta de algunas funcionalidades básicas como la impresión de cartografía, cuadrículas, etc. Muchas de estas funcionalidades están en vías de solución.
- Problemas con la coordinación en la generación de versiones, aunque actualmente se ha creado un comité para coordinar el desarrollo de las futuras versiones.

2.4.6. uDIG

Se puede considerar uDIG como el sucesor de Open Jump en muchos aspectos. Conceptualmente uDIG utiliza OpenJump como base de algoritmos para el manejo y manipulación de datos espaciales y Geotools como librería para la entrada y salida de datos, con lo que se asegura un buen número de formatos soportados (Shekhar y Xiong, 2008).

uDIG tiene su origen en la empresa Refractions (creadores de PostGIS), y uno de sus objetivos es basarse firmemente en estándares del OGC (Ramsey, 2003). Está programado en Java y aunque actualmente se encuentra en una fase inicial de desarrollo, por la evolución y las declaraciones de intenciones del proyecto parece ser muy prometedor. El punto más importante a destacar es que permite la conexión a servidores WFS en modo lectura y escritura, algo que no soportan muchos de los sistemas de información geográfica libres.

Sus principales puntos fuertes son:

- Interfaz de usuario muy intuitiva.
- Buen número de formatos soportados a través de *plugins*, incluyendo conexión a servidores.
- Buen punto de partida para la creación de proyectos personalizados debido a la documentación existente y a la facilidad de implementación de nuevas funciones.

Presenta algunas dificultades, como por ejemplo:

- Soporte del estándar WFS tanto en lectura como escritura.

- Soporte para servidores WMS.
- Soporte de acceso a todos los datos soportados por Geotools, tanto de archivos como de servidores de bases de datos PostGIS o MySQL.
- Capacidad de impresión y salidas gráficas en diversos formatos.
- Diseño modular orientado a la reutilización en otros proyecto o programas.

Actualmente uDIG no posee muchas opciones de visualización y edición.

2.4.7. Sextante

Sextante es un conjunto de algoritmos de análisis geoespacial de código libre desarrollado para la Junta de Extremadura, al cual se le pueden implementar y añadir nuevos algoritmos; este desarrollo se lleva a cabo tomando como base otro *software* existente, e implementándole un grupo de nuevas capacidades (Olaya, 2008).

En la actualidad Sextante está integrado en algunos de los sistemas de información geográfica más populares escritos en Java y también se puede acoplar a otros no desarrollados en este lenguaje de programación, lo cual muestra su aceptación dentro de las comunidades de sistemas de información geográfica. Sextante inicialmente tuvo como base el sistema de información geográfica alemán Saga, para el cual se desarrolló una gran cantidad de extensiones y modificaciones en su núcleo base (Olaya, 2011a). Desde la versión 1.10 de gvSIG este ha sustituido a Saga como software base, principalmente por contar con una estructura de apoyo más sólida y con un mayor potencial futuro.

2.5. Análisis exploratorio de datos. Fundamentos

El análisis de datos en la estadística se puede entender como "el proceso de calcular varios valores derivados o resumidos dada una colección de datos" (Hand, 1999). Hay que enfatizar que este proceso es iterativo: "Quien estudia los datos, los examina utilizando alguna técnica analítica, decide buscar de otra forma, tal vez modificándolos en el proceso por alguna transformación o partición, y luego regresa al inicio y aplica otra herramienta de análisis de datos. Esto puede iterar varias veces. Cada técnica es utilizada para sondear ligeramente aspectos de los datos para preguntar interrogantes diferentes sobre los datos" (Hand, 1999).

En el área de los sistemas de información geográfica, el análisis de datos es a menudo definido como: "un proceso de búsqueda de patrones geográficos en los datos y en las relaciones entre objetos geográficos" (Mitchell, 1999). Comienza con la formulación de preguntas que necesitan ser contestadas, luego se escoge un método básico sobre la base de la pregunta, el tipo de dato disponible y el nivel de información requerido (esto puede aumentar la necesidad de utilizar datos adicionales). Luego los datos son procesados mediante el método seleccionado y

los resultados son visualizados, lo que permite a los analistas decidir si la información obtenida es válida o útil, y en función de esto realizar de nuevo el análisis con la utilización de diferentes parámetros o incluso un método diferente.

Lo que es común para estas dos definiciones es que ven el análisis de datos como un proceso iterativo que consiste de las actividades siguientes:

- Formular preguntas.
- Seleccionar un método de análisis.
- Preparar los datos para la aplicación de los métodos.
- Aplicar el método a los datos.
- Interpretar y evaluar los resultados obtenidos.

La diferencia entre el análisis estadístico y el análisis con sistemas de información geográfica radica solo en los tipos de datos con que tiene que ver cada tipo de análisis y con el método utilizado. En ambos casos, el análisis de datos está dirigido por preguntas: las preguntas motivan la realización del análisis, determinan la selección de los datos y métodos, y afectan la interpretación de los resultados. Es por eso que las preguntas son tan importantes.

Ni los libros de estadística ni los de sistemas de información geográfica suministran una clasificación para posibles preguntas, pero en su lugar ellos suministran algunos ejemplos como los que se pueden ver en Mitchell (1999) y en Burt y Barber (1996).

La familiarización con los datos es el tema seguido en el análisis exploratorio de datos, aunque a veces solo se quiere explorar un conjunto de datos para ver qué nos puede decir. Cuando se hace esto, se está haciendo análisis exploratorio de datos (Wildman, 2005).

Aunque el análisis exploratorio de datos surgió de la estadística, este no es un conjunto específico de técnicas a diferencia de la estadística en sí, sino una filosofía de cómo el análisis de datos puede llevarse a cabo. Esta filosofía fue definida por Tukey (1977) para contrarrestar el sesgo que existía en la investigación estadística con el desarrollo de métodos matemáticos de prueba de hipótesis. Según lo valoró Tukey, el análisis exploratorio de datos fue el regreso a los objetivos originales de la estadística, por ejemplo detectar y describir patrones, tendencias, y relaciones entre los datos. En otras palabras, el análisis exploratorio de datos está más relacionado con la generación de hipótesis que con la prueba de hipótesis (Andrienko y Andrienko, 2006).

El concepto análisis exploratorio de datos está fuertemente asociado con el uso de representaciones gráficas de los datos. De ahí la relación de este concepto con la visualización presentada en epígrafes anteriores. La mayoría de los métodos de análisis exploratorio de datos son de naturaleza gráfica con unas pocas técnicas cuantitativas. La razón para confiar tanto en los gráficos está dada por el rol principal del análisis exploratorio de datos, explorar con mente abierta los gráficos, explotando el poder humano para interpretar imágenes y reconocer patrones, lograr que los datos revelen sus secretos estructurales y estar siempre listos para ganar nuevo conocimiento sobre los datos (Andrienko y Andrienko, 2006).

Según la bien conocida *Information Seeking Mantra*, introducida por Shneiderman (1996), el análisis exploratorio de datos puede ser generalizado como un proceso de 3 pasos: "obtener una visión general primero, filtrar y ampliar en algunos aspectos, y luego buscar detalles a demanda". En el primer paso, un analista necesita obtener una visión general de la colección completa de datos. Aquí el analista identifica elementos de interés. En el segundo paso, amplía la búsqueda en los elementos de interés y desecha los elementos no interesantes. En el tercer paso, selecciona un elemento o grupo de elementos para profundizar y obtener detalles. Nuevamente, el proceso es iterativo con varios regresos al paso anterior.

Sobre esta base Andrienko y Andrienko (2006), adoptan la siguiente visión del análisis exploratorio de datos. El analista tiene un propósito en su investigación, que motiva el análisis. El propósito se especifica como una pregunta general o un conjunto de preguntas generales. El analista comienza el análisis buscando qué es interesante en los datos, donde la condición de ser interesante se entiende como relevancia para el propósito de la investigación. Cuando se detecta algo interesante, aparecen nuevas preguntas más específicas que motivan a los analistas a buscar los detalles. Estas preguntas afectan los detalles que van a ser observados y de qué forma. Es por eso que las preguntas juegan un rol importante en el análisis exploratorio de datos y pueden determinar la selección del método de análisis (Andrienko y Andrienko, 2006).

En los libros de texto de estadística y de sistemas de información geográfica existen pocas distinciones en comparación con las preguntas dadas:

- El análisis exploratorio de datos esencialmente involucra muchas preguntas diferentes.
- Las preguntas varían con el nivel de generalidad.
- La mayoría de las preguntas principalmente se originan durante el análisis, raras veces se tienen preconcebidas.

Existe una coincidencia en la multitud y diversidad de preguntas involucradas en el análisis exploratorio de datos: este tipo de análisis requiere el uso en combinación de múltiples herramientas y técnicas, ya que ninguna herramienta simple puede suministrar respuestas a todas las preguntas. Un sistema ideal que pretenda soportar análisis exploratorio de datos, debe contener un conjunto de herramientas que puedan ayudar a los analistas a responder cualquier pregunta posible (por supuesto, sólo si dispone de la información necesaria). Este ideal probablemente nunca pueda lograrse, pero un diseñador debe concebir un sistema o *kit* de herramientas para el análisis de datos y necesita anticipar las potenciales preguntas o al menos hacer una selección de cuáles debe soportar.

En este epígrafe se describen elementos esenciales que se deben tener en cuenta en el análisis exploratorio de datos desde diferentes perspectivas.

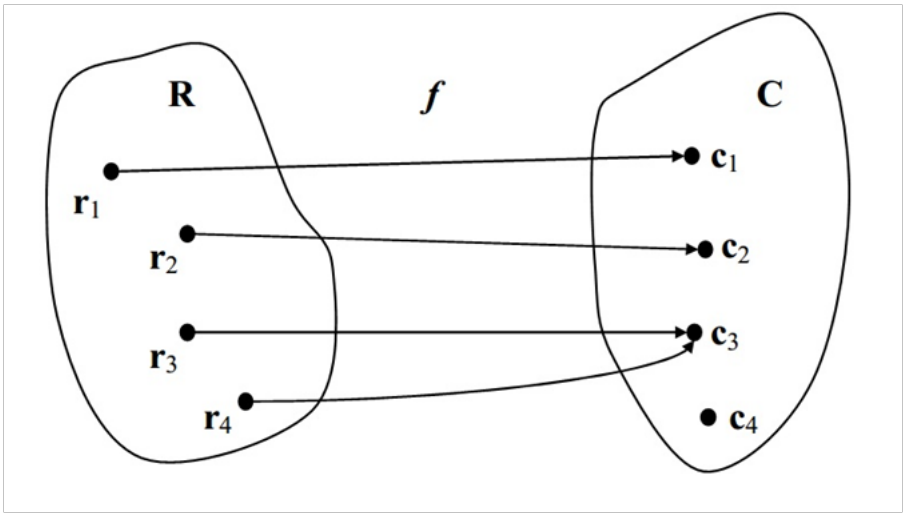

Figura 2.10 Vista funcional de un conjunto de datos. Tomado de (Andrienko y Andrienko, 2006)

2.5.1. Datos

El punto de partida del análisis exploratorio de datos lo constituyen los datos. Los datos describen hechos de una manera estructurada, semi-estructurada, o sin ningún tipo de estructura, y deben ser representativos y estar relacionados con el problema analítico; de otra forma es improbable que se puedan descubrir relaciones significativas en el dominio del problema.

El análisis exploratorio de datos puede analizarse desde varios puntos de vista, como proponen Andrienko y Andrienko (2006), desde el punto de vista de los datos, tareas, herramientas y principios.

Desde la perspectiva de los datos el punto esencial es distinguir entre los componentes de datos características y referencias: las características reflejan observaciones o mediciones, mientras que las referencias especifican el contexto en que estas observaciones o medidas fueron tomadas, por ejemplo lugar y/o tiempo.

En esta tesis se comparte la idea de Andrienko y Andrienko (2006), de ver un conjunto de datos como una función (en el sentido matemático), donde se establecen enlaces entre referencias (por ejemplo, indicadores particulares de lugar, tiempo, etc.) y características (por ejemplo, mediciones particulares o valores observados). Esta función puede ser representada simbólicamente como muestra la figura 2.10, donde r_1, r_2, r_3 y r_4, representan diferentes referencias, por ejemplo combinaciones de valores del conjunto de datos referencia. R es el conjunto de todas las referencias r_i. c_1, c_2, c_3 y c_4 representan diferentes características, por ejemplo, combinaciones de valores de atributos. C es el conjunto de todas las posibles características c_i. f es la función de datos, que asocia cada referencia con su correspondiente característica.

Andrienko y Andrienko (2006) muestran este concepto mediante siete ejemplos específicos de conjuntos de datos. La figura 2.11 presenta la modificación de uno de los ejemplos. En este caso las referencias corresponden a celdas del espacio y los momentos de tiempo; las características están dadas por atributos que contienen información para cada uno de los elementos

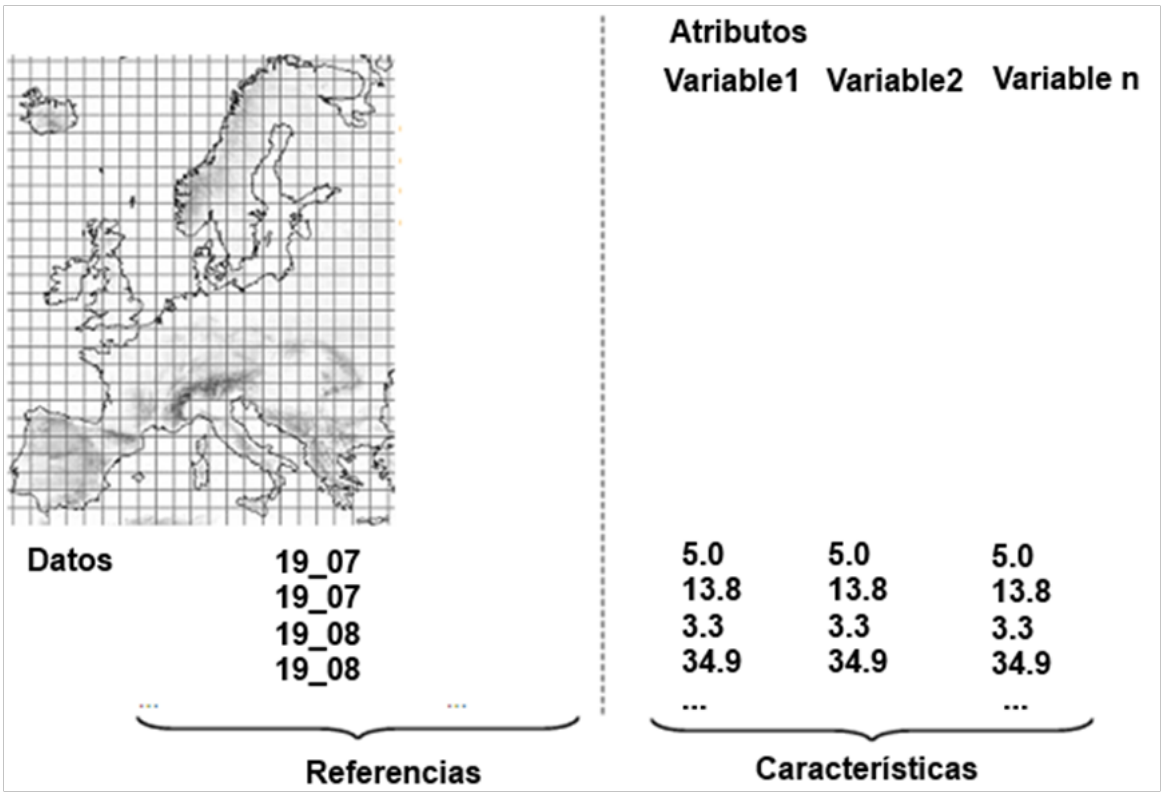

Figura 2.11 Representación visual de la estructura de un conjunto de datos. Modificado de (Andrienko y Andrienko, 2006)

referencia.

2.5.2. Tareas

Las tareas, es decir, las preguntas que se necesitan responder por medio del análisis de datos se definen en términos de los componentes de datos. Como se puede ver en la figura 2.12, ambos gráficos representan esquemáticamente una tarea. ¿Cuáles son las características que corresponden a una referencia dada?, y ¿cuál es la referencia correspondiente a una característica dada?

Andrienko y Andrienko (2006), plantean que un punto esencial es la distinción entre tareas elementales o sinópticas. *Elemental* no significa simple, aunque generalmente suelen ser más simples que las sinópticas. Las tareas elementales son las que tienen que ver con elementos de datos, es decir con referencias y características individuales. Las tareas sinópticas tienen que ver con conjuntos de referencias y sus correspondientes configuraciones de características. Ambas se consideran como un todo unificado. Esta clasificación de tareas en elementales y sinópticas proviene de las ideas de Bertin (1983).

El término *comportamiento* y *patrón* se define en Andrienko y Andrienko (2006). *Comportamiento* denota una configuración particular de características existente. El término de *patrón* denota la forma en que se observa e interpreta un comportamiento, y cómo este se presenta a

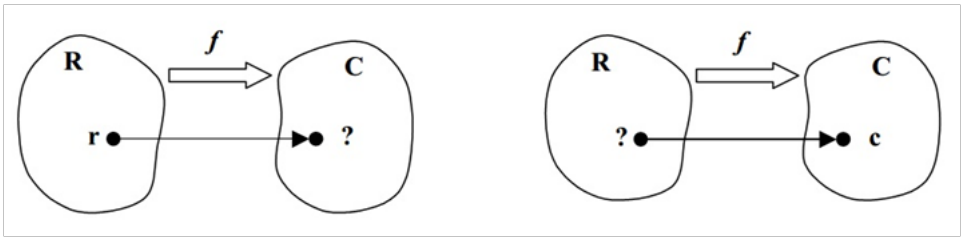

Figura 2.12 Definición de dos tipos de tareas representadas sobre la base de la visión funcional de datos. Tomado de (Andrienko y Andrienko, 2006)

otras personas. Por ejemplo, se puede calificar el comportamiento de la temperatura del aire del mediodía durante la primera semana de abril, como con una tendencia a incrementarse: *tendencia a incrementarse* es el patrón resultante desde esta apreciación de comportamiento.

El principal objetivo del análisis exploratorio de datos puede ser visto generalmente como la construcción de un patrón apropiado del comportamiento general definido por el conjunto de datos completo (Andrienko y Andrienko, 2006). Por ejemplo, ¿cuál es el comportamiento de la estructura de los bosques en el territorio de Europa?, o ¿cuál es el comportamiento del clima de Alemania durante el período de 1991 a 2003?

Cuando se exploran datos multidimensionales se pueden tener dos o más componentes referenciales, por ejemplo el espacio y el tiempo. El comportamiento de datos multidimensionales puede ser visto desde diferentes perspectivas y cada perspectiva puede revelar algunos aspectos, a lo que se le puede llamar *comportamiento aspectual* (Andrienko y Andrienko, 2006). En principio cada comportamiento aspectual necesita ser analizado, pero el número de estos comportamientos aumenta rápidamente con el incremento del número de componentes referenciales.

Las tareas elementales pueden ser de varios tipos: tareas de búsqueda (directa o inversa), tareas de comparación (directa o inversa) y tareas de búsqueda de relaciones. Las tareas sinópticas pueden incluir estos tipos: Caracterización de comportamientos, búsqueda de patrones de comportamientos, comparación de comportamientos, y descubrimiento de conexiones entre comportamientos.

La taxonomía de tareas por tipo de datos brindada por Shneiderman (1996), enumera siete tareas de alto nivel, que también incluyen la noción de interacción con los datos, además de las tareas puramente visuales (Plaisant, 2005; Aigner, 2011):

- Descripción general: obtener una visión general de todo el conjunto de datos.
- *Zoom*: enfocar los datos de interés.
- Filtro: filtrar la información relevante.
- Detalles a demanda: cuando es necesario seleccionar los datos de interés y obtener detalles.
- Relacionar: ver las relaciones entre los elementos de datos.
- Historia: mantener un historial de acciones de apoyo a deshacer y rehacer.
- Extraer: permite la extracción de los datos y de consulta de parámetros.

Yi *et al.* (2007) redefinieron el aspecto de la interacción en la visualización, del cual se derivó una serie de categorías de tareas de interacción. Estas categorías están relacionadas con las intenciones del usuario para ajustar representaciones visuales a tareas y datos, de manera interactiva. En consecuencia, la tarea "muestra" tiene seis categorías:

- muestra algo más (explorar)
- muestra una organización diferente (reconfigurar)
- muestra una representación diferente (codificar)
- muestra más o menos detalle (resumir / elaborar)
- muestra algo condicional (filtrar)

Las tareas "muestra" permiten cambiar entre diferentes subgrupos de datos analizados (explorar), diferentes organizaciones de primitivas visuales (reconfigurar), y diferentes representaciones visuales (codificar). También trata la navegación entre los diferentes niveles de detalle (resumir / elaborar), la definición de los datos de interés (filtrar), y la exploración de las relaciones (conectar). Además de las tareas "muestra", Yi *et al.* (2007) introducen tres tareas de interacción adicionales:

- Marcar algo como interesante (seleccionar)
- Permitir volver a donde ya ha ido (deshacer / rehacer)
- Permitir ajustar la interfaz (cambiar configuración)

Marcar algo como interesante (seleccionar), resume todo tipo de tareas de selección, incluyendo tanto la selección de valores de datos individuales, como la selección de subconjuntos completos. Permitir a los usuarios regresar a datos interesantes o vistas (deshacer / rehacer), es esencial durante la exploración interactiva de datos. La adaptabilidad (cambiar la configuración) es relevante cuando un sistema se utiliza por una amplia gama de usuarios, para una variedad de tareas y tipos de datos.

2.5.3. Herramientas

Las herramientas pueden ser divididas en cinco categorías: herramientas de visualización, manipulación de vistas, manipulación de datos, de consulta, y de cálculo (Andrienko y Andrienko, 2006).

La visualización como categoría dentro de las herramientas es la representación de datos de una manera visual, al crear varias imágenes con los datos: gráficos, diagramas, mapas, etc. Para esto los elementos de datos son transformados en características gráficas, tales como: posición dentro de la pantalla, colores, tamaños, o formas. Es importante mencionar, sin embargo, que

estas características gráficas se combinan en una simple imagen, observándose como un todo en lugar de ser percibidas separadamente.

La manipulación de vistas consiste en herramientas interactivas que soportan la modificación dinámica de la apariencia visual de las vistas. El propósito general de tal modificación consiste en mejorar la imagen producida para hacerla más clara y fácil de percibir, resaltar las características distintivas presentadas, enfocar en un elemento particular o subconjunto de interés, etc. La manipulación se realiza a través de la modificación de fórmulas o algoritmos utilizados para la transformación de los elementos de datos en características visuales, a esto se le conoce como *la función de codificación visual*.

La manipulación de datos es la derivación de nuevas referencias y características a partir de las existentes. Esto se hace con dos propósitos principales: simplificar los datos y hacerlos fácil de analizar o, por el contrario, enriquecer y considerar varios aspectos de ellos. De esta manera la agregación de datos reduce la cantidad de datos y, por lo tanto, simplifica el análisis. En contraste la interpolación produce datos adicionales.

La consulta es la búsqueda automática de respuestas a preguntas especificadas por el usuario. Esto es la búsqueda de referencias con características especificadas o buscar características de referencias especificadas. Para el análisis exploratorio de datos son especialmente importantes las herramientas de consultas dinámicas, las cuales permiten al usuario modificar fácilmente las condiciones de consultas y suministrar rápidamente la respuesta requerida.

Entre herramientas de cálculo se consideran los métodos computacionales de la estadística y la minería de datos. A diferencia de los cálculos involucrados en la manipulación de datos, que preparan los datos para análisis futuros, por ejemplo, transformando los datos de una forma más adecuada, la función de las herramientas computacionales es un tipo de limpieza de datos o extracción de las características esenciales en los datos. Algunos ejemplos de las salidas producidas por las herramientas computacionales son características estadísticas de un conjunto de datos como un todo, indicadores de relevancia entre atributos y modelos que predigan algunas características sobre la base de otras características, en particular desarrollos futuros sobre la base del estado actual y de la historia.

En el análisis exploratorio de datos no es suficiente el uso de una simple herramienta; se necesita la combinación de varias herramientas, donde la visualización es un componente esencial. La visualización de datos en un inicio se utiliza para comprender qué herramientas deberán ser utilizadas en trabajos futuros. Los resultados producidos por cualquier herramienta no visual necesitan ser visualizados para que un analista los pueda ver e interpretar. En los epígrafes 2.1, 2.2 y 2.3 se presentaron elementos sobre la visualización, no obstante en capítulos posteriores se mostrará la utilización de otros tipos de herramienta que son también importantes para el análisis exploratorio de datos.

2.5.4. Principios

Como se ha mencionado anteriormente, la exploración visual de datos por lo general sigue un proceso de tres etapas: primero obtener una visión general (del término en inglés *Overview first*), luego ampliar/reducir y filtrar (zoom del término en inglés *zoom and filter*), y por último detallar por demanda (del término en inglés *details-on-demands*). A esto se le ha denominado *Information Seeking Mantra* (Shneiderman, 1996). Primeramente, el usuario debe tener una visión general de los datos. En este paso, el usuario identifica patrones interesantes o agrupaciones en los datos y se centra en uno de ellos o más. Para el análisis de los patrones, el usuario necesita indagar en los datos y acceder a los detalles. Las tecnologías de la visualización pueden ser usadas para cualquiera de los tres pasos en el proceso de exploración de datos. Las técnicas de visualización son útiles para mostrar visiones generales y para permitir al usuario identificar subgrupos de interés. En este paso, es importante mantener una visualización general mientras se concentra en un subgrupo mediante otra técnica de visualización.

Existe un grupo de principios generales que puede ayudar en la selección de herramientas para la exploración de datos (Andrienko y Andrienko, 2006):

- Ver el todo. Este principio permite representar los datos de tal manera que se pueda percibir el comportamiento general por medio de la visión. Esto requiere, primeramente, que no se omita algo esencial (idealmente, que se presenten en una vista todos los elementos de datos), en segundo lugar que todos los aspectos sean reflejados, y por último que los elementos visuales que representan los datos sean percibidos de una vez como un todo unificado.
- Simplificar y resumir. Este principio permite descartar los detalles excesivos y peculiaridades ocasionales que obstruyan la percepción de las características esenciales del comportamiento de los datos.
- Dividir y agrupar. Cuando se ve o se espera que el comportamiento general no sea el mismo en todo el conjunto referencia, se divide este en conjuntos de tal forma que el comportamiento dentro de cada subconjunto pueda ser suficientemente homogéneo. Entonces, el comportamiento general puede ser caracterizado como una combinación del comportamiento parcial.
- Ver relaciones. Para una adecuada caracterización del comportamiento dividido en partes, se deben revelar las diferencias substanciales así como las similitudes entre las partes. Es también importante comparar el comportamiento de los diferentes atributos o grupos de atributos.
- Buscar lo reconocible. Representa los datos de manera tal que puedan ser fácilmente detectados ordenamientos específicos de características o subpatrones. La característica a buscar depende de la estructura y naturaleza de los datos.
- Ampliar y enfocar. En la exploración del comportamiento parcial sobre subconjuntos referencia, aplicar herramientas que ayuden a concentrarse en la parte que actualmente

se está analizando y representar esta parte con el máximo posible de expresividad. Sin embargo, es importante posicionar esta parte con respecto al comportamiento completo, por ejemplo ver el contexto.

- Atender a particulares. Detecta, examina a fondo y trata de explicar varios casos de características inusuales.
- Establecer enlaces. Integra en una vista coherente la observación y los patrones parciales derivados de la investigación de varias partes y aspectos del comportamiento general.
- Establecer estructura. Cuando se sospecha que el comportamiento general resulta de la interacción de varios componentes estructurales, tales como procesos lineales y cíclicos en fenómenos relacionados con el tiempo, hay que explorar cada componente y sus interacciones con otros mediante la división de referencias relevantes en varias referencias, o introduciendo referencias adicionales.
- Involucrar conocimiento sobre el dominio. Siempre que sea posible, hacer uso sobre lo que se conoce acerca de la naturaleza y propiedades de los fenómenos relacionados con los datos, o incluso hacer uso del sentido común. Esto puede hacer que ocurra una anticipación a las tendencias generales en el comportamiento, o distinciones importantes entre ciertas partes de los datos, del ordenamiento de características (subpatrones).

2.6. Integración de visualización científica con sistemas de información geográfica para el análisis exploratorio de datos

Por más de dos décadas la visualización científica y los sistemas de información geográfica se desarrollaron en paralelo y de forma independiente (Rhyne, 1997). Los esfuerzos para desarrollar estándares de datos espaciales raras veces consideraron la forma en que estos se visualizaban. Las bibliotecas gráficas y los estándares evolucionaron independientemente de los modelos de datos. Como resultado de esto se evidenciaron muchas ineficiencias asociadas con la visualización de datos geográficos. Entre ellas, se incluyen: dificultades con el registro de los datos espaciales dentro de sistemas de visualización, engorrosas producciones de secuencias de animaciones en sistemas de información geográfica y, quizás la más importante, la falta de conexión entre bases de datos y los ambientes de visualización que soportaban la visualización de datos espaciales (Hearnshaw y Unwin, 1994).

Los desarrolladores de herramientas de sistemas de información geográfica y de visualización científica hicieron esfuerzos para ampliar e integrar sus sistemas (Rhyne *et al.*, 1994). Los desarrolladores de sistemas de información geográfica estudiaron la forma de incorporar las capacidades de la animación de series de tiempo en tres dimensiones en su software. Los desarrolladores de herramientas de visualización científica comenzaron la construcción de lectores de datos que soportaban los formatos de datos espaciales como modelos digitales de elevación,

así como formatos de sistemas de información geográfica comerciales.

Al examinar estos esfuerzos fueron definidos cuatro niveles de métodos de integración entre sistemas de información geográfica y visualización científica: rudimentario, operacional, funcional y mezclado (Rhyne, 1997).

El enfoque rudimentario utiliza una mínima integración de datos e intercambio entre las dos tecnologías. El nivel operacional proporciona coherencia entre los datos mientras se eliminan las redundancias entre las dos tecnologías (Cook *et al.*, 1997). La forma funcional intenta proporcionar una comunicación transparente entre los entornos de software correspondientes (Mitas *et al.*, 1997). El enfoque mezclado se refiere al desarrollo de sistemas donde los conceptos de cartografía, sistemas de información geográfica y visualización científica se funden en una única herramienta.

En casos puntuales se ha logrado el nivel rudimentario de intercambiar datos en formatos de sistemas de información geográfica hacia las herramientas de visualización científica. Algunos ambientes de visualización científica se han aproximado en el nivel operacional, y permitido accesos directos a bases de datos de sistemas de información geográfica; sin embargo, esto suele ser en un solo sentido, una vez que la herramienta de visualización científica genera la imagen tridimensional o la animación, generalmente no es posible activar las funciones de consulta de los sistemas de información geográfica desde la pantalla de visualización.

Lograr la integración funcional de sistemas de información geográfica y herramientas de visualización científica requiere de estándares abiertos de datos de sistemas de información geográfica, enlaces a programas que permitan que las herramientas de visualización científica realicen análisis de datos espaciales y funciones de extracción de información. También es factible el uso de sistemas expertos o una arquitectura basada en reglas con agentes inteligentes para facilitar la comunicación transparente entre sistemas de información geográfica y herramientas de visualización científica (Rogowitz y Treinish, 1993).

Los sistemas que implementan el enfoque mezclado replantean el proceso de desarrollo de herramientas de sistemas de información geográfica y visualización científica. En este sentido la cartografía está bien posicionada como puente entre ambas tecnologías.

Algunos de los primeros intentos de integración entre los sistemas de información geográfica y la visualización científica se han materializado a través de herramientas como GeoVista Studio (Gahegan *et al.*, 2002; Takatsuka y Gahegan, 2002; MacEachren *et al.*, 2003) y Snap-Together Visualization (North y Shneiderman, 2000). La mayoría de los paquetes que se dedican a la visualización de datos espacio-temporales se enfocan más hacia las visualizaciones que hacia la integración con el sistema de información geográfica. Utilizan los sistemas de información geográfica de manera separada para preparar datos, exportar e importar datos y finalmente hacer las visualizaciones con una herramienta determinada. VIS-STAMP (Guo *et al.*, 2006) y GAV Flash (Ho *et al.*, 2011) son ejemplos de integración más recientes. Estas herramientas son descritas en este epígrafe.

De igual manera, se han desarrollado intentos de integrar algunos de los formatos de datos

científicos más comúnmente utilizados en la visualización científica –como *Common Data Format* (CDF), *Network Common Data Format* (NetCDF), *Hierarchical Data Format* (HDF), HDF-EOS y *Flexible Image Transport System* (FITS)–, con sistemas de información geográfica, por ejemplo actualmente la suite de ArcGIS permite la manipulación de algunos de estos formatos, como NetCDF y HDF (Zhao *et al.*, 2010). El epígrafe 2.7 describe estos formatos de datos científicos.

La visualización geográfica o geovisualización, es un campo emergente que se destaca por la incorporación de técnicas y herramientas para el análisis visual interactivo de datos espaciales y espacio-temporales. Una característica distintiva es la integración de los enfoques de múltiples disciplinas como la geografía, la ciencia de información geográfica, cartografía, visualización de información, minería de datos y otras disciplinas afines (Keim *et al.*, 2010).

En este epígrafe se describen algunas de las principales herramientas que permiten hacer geovisualización exploratoria de datos.

2.6.1. Snap-Together Visualization

Snap-Together Visualization es una herramienta Web que permite que los datos de los usuarios sean mezclados y correlacionados de forma dinámica en visualizaciones coordinadas, para la construcción personalizada de interfaces de exploración sin necesidad de programación. El modelo conceptual de Snap-Together se basa en un modelo de base de datos relacional. Este modelo permite que las relaciones sean cargadas durante la visualización y se coordinen basándose en las características que las unen. Los usuarios pueden crear diferentes tipos de coordinaciones tales como: barridos, vistas de detalles, vistas globales y desplazamientos sincronizados.

Los desarrolladores de esta herramienta de visualización pueden integrar al sistema sus visualizaciones independientes con una API simple. La evaluación de esta herramienta reveló beneficios en cuanto a aspectos cognitivos y su usabilidad, mejorando el rendimiento de los usuarios entre un 30 y un 80 por ciento, en dependencia de la tarea realizada. Algunas de las técnicas de visualización que están incluidas en esa herramienta son Diagramas de Dispersión y TreeMap (North y Shneiderman, 2000).

Snap-Together Visualization es una herramienta orientada a la Web, tiene la desventaja de que no es de código abierto y no soporta la manipulación de grandes volúmenes de datos en forma de mallas regulares; elementos que dificultan su accesibilidad y su usabilidad.

2.6.2. Geovista Studio

GeoVista Studio (Gahegan *et al.*, 2002; Takatsuka y Gahegan, 2002; MacEachren *et al.*, 2003) es una herramienta de código abierto que implementa un ambiente de desarrollo basado

en componentes. Suministra, como muchos sistemas de visualización científica, una interfaz de programación visual a través de la cual los usuarios pueden construir aplicaciones de forma rápida utilizando JavaBeans. El ambiente de programación visual permite a los analistas empaquetar funcionalidades dentro de un programa de trabajo. Geovista Studio soporta tanto el desarrollo de aplicaciones geográficas como no geográficas (Luo *et al.*, 2014).

Para soportar la interoperabilidad de datos OpenGIS, sus desarrolladores han comenzado a adaptar y extender la biblioteca GeoTools en cuanto al acceso a datos y los métodos de visualización. GeoTools es una biblioteca complementaria de software libre desarrollada en Java para el desarrollo de soluciones OpenGIS que permite el acceso a datos geoespaciales, el análisis y la representación de tareas (Turton, 2008).

El principal objetivo de GeoVista es soportar la fusión de diversas capacidades visuales y analíticas en una herramienta de análisis que posibilite la multiperspectiva. Incluye además un conjunto de técnicas de visualización clásicas como diagramas de dispersión, coordenadas paralelas y mapas auto-organizados (Kohonen, 1990; Gahegan *et al.*, 2002; Takatsuka y Gahegan, 2002). Varias herramientas han sido desarrolladas tomando como base el Geovista Studio, entre ellas se encuentran el *Exploratory Spatio-Temporal Analysis Toolkit ESTAT* (Robinson *et al.*, 2005) y VIS-STAMP (Zhang *et al.*, 2013).

Geovista posibilita la utilización de formatos de datos vectoriales clásicos de los sistemas de información geográfica como el formato *shp*, sin embargo, no posee herramientas para manipular formatos *raster* o formatos de datos científicos. A pesar de que provee herramientas para analizar series temporales, los datos de series temporales son manipulados a partir de archivos de texto plano o separados por coma *csv*. Cuando aumenta la cantidad de variables de las series temporales, el sistema puede ver afectado su rendimiento con algunas de las visualizaciones.

2.6.3. VIS-STAMP

VIS-STAMP es un paquete de software que integra métodos computacionales, visuales y cartográficos para la exploración y visualización de datos espacio-temporales multivariados. Uno de sus principales objetivos es proveer los medios para el descubrimiento de patrones desconocidos (Guo *et al.*, 2006). En particular ha sido usado exitosamente en el estudio de patrones relacionados con el cambio climático (Jin y Guo, 2009). La base del sistema es un mapa auto-organizado usado para el agrupamiento, ordenamiento y coloración (Aigner, 2011).

El sistema está diseñado para ser flexible y soporta el uso de diferentes métodos de búsqueda de conglomerados. La presentación visual de los datos está compuesta por varias vistas integradas (interrelacionadas entre sí). En particular incluye mapas geospaciales, mapas auto-organizados, gráficos de malla y coordenadas paralelas (Andrienko *et al.*, 2010b).

VIS-STAMP posibilita la utilización de formatos de datos vectoriales clásicos de los sistemas de información geográfica como el formato *shp*. No soporta la manipulación de formatos de datos científicos. Posee pocas técnicas de visualización para realizar análisis visuales. Puede ver

afectado su rendimiento cuando el volumen de datos es grande.

2.6.4. GAV Flash tools

Las herramientas GAV Flash (Andrienko *et al.*, 2010a; Ho *et al.*, 2011) del término en inglés *Geovisual Analytics Visualization* comparten aplicaciones basadas en los principios del análisis visual de datos. Contienen una colección de componentes visuales, algoritmos de análisis de datos, herramientas que conectan los componentes con otros componentes y suministradores de datos que pueden cargar datos desde varias fuentes. El sistema está completamente integrado con el framework de Adobe Flex (Ho *et al.*, 2012).

GAV Flash ha sido programado con programación orientada a objetos en el lenguaje Action Script, facilita un 100 por ciento de despliegue en Internet a través de Adobe Flash Player versión 10 (Jern *et al.*, 2009). Es una herramienta interactiva que apoya el proceso de razonamiento analítico espacial, provee herramientas para buscar, filtrar y resaltar datos mediante consultas, puede ser utilizado para descubrir valores atípicos o fuera de rango. Además, implementa métodos para enlazar múltiples vistas.

Como GAV Flash se basa en Adobe Flex, un desarrollador tiene acceso a todas las funcionalidades de la interfaz de usuario de Flex. Mediante la combinación de botones, paneles y deslizadores, con los proveedores de datos de GAV Flash se pueden configurar manipuladores y representaciones visuales de una manera sencilla. La arquitectura abierta que posee, permite la incorporación de nuevas herramientas o componentes existentes, como por ejemplo, herramientas de análisis estadístico o técnicas de visualización. La separación de las estructuras de datos de las técnicas de visualización, permite crear aplicaciones que trabajan independientemente de la entrada, por lo que los datos pueden ser suministrados desde el exterior y ser conectados con el sistema con un mínimo de conocimientos de programación.

GAV Flash posee algunas desventajas relacionadas con la accesibilidad y la usabilidad, a pesar de poseer una arquitectura abierta, no se dispone gratuitamente del código fuente. El número de técnicas de visualización que posee actualmente no se considera alto. Además, no manipula formatos de datos *raster* ni formatos de datos científicos, lo que le dificulta el tratamiento de grandes volúmenes de datos.

2.6.5. ArcView-xGobi

En la geovisualzación, un buen ejemplo de integración es el enlace bidireccional entre ArcView y XGobi (Symanzik *et al.*, 2000), trabajo donde se integran los gráficos interactivos de XGobi para manipular datos con muchas dimensiones con las herramientas de manipulación de datos espaciales de ArcView. La integración *ArcView - XGobi - Xplore*, permite que los datos

recogidos en lugares espaciales que se almacenan en ArcView, pasen dinámicamente a XGobi y Xplore y puedan ser explorados y analizados.

El vínculo entre los datos de XGobi y Xplore y los lugares de los que fueron recogidos, se mantienen a través de marcado y enlace (*linking and brushing*). El marcado y enlazado, tal como se utiliza en este contexto, es la capacidad de cambiar el tamaño / color de los puntos, ya sea en ArcView, XGobi o Xplore. De esta forma se puede notar que los puntos correspondientes de las otras aplicaciones cambian simultáneamente.

Actualmente esta herramienta ha quedado obsoleta. No se han desarrollado nuevos enlaces de XGobi con las versiones actuales de ArcGIS. La principal desventaja es que ArcView no es de código abierto y se requiere el pago de licencias.

2.7. Formatos de datos científicos espacio-temporales

Los sistemas de información geográfica facilitan la gestión y representación de datos espaciales, y permiten modelar el comportamiento de diversas situaciones en el contexto espacial (Murphy, 1995). Varios estudios han permitido desarrollar herramientas que posibilitan el análisis de variables temporales. Estos estudios se fundamentan en el hecho de que los datos espaciales son una representación del mundo real y por tanto esto implica la incorporación de modelos dinámicos dependientes de variables espaciales y temporales (Vanegas, 2013). El tiempo es una dimensión fundamental para entender la ocurrencia de los fenómenos geográficos. Además, estos datos tienen gran aplicación para el desarrollo de modelos terrestres y de predicción del cambio climático global, por lo cual son muy importantes para la toma de decisiones sobre la protección del medio ambiente (Esparza Gil, 2014).

Una entidad geográfica tiene una ruta espacio-temporal desde que inicia el momento de su creación hasta que termina en el momento que se destruye. Durante esta etapa su ciclo de vida sufre varios cambios en su localización y también se ve afectado por eventos que ocurren en el tiempo. Por lo que se puede afirmar que el espacio y el tiempo son prácticamente inseparables (Jaramillo y Vanegas, 2013).

Los datos espacio-temporales se pueden representar como una inserción de la dimensión tiempo en entidades geográficas concebidas, donde el espacio geográfico se organiza en capas temáticas que incluyen la información de captura en un tiempo determinado (Rodríguez *et al.*, 2009).

Existen diversos formatos de datos científicos espacio-temporales, que son usados por diferentes comunidades e instituciones científicas para almacenar e intercambiar grandes volúmenes de información (McGrath, 2003). Pero estos formatos tienen diferentes niveles de uso de acuerdo a las comunidades que los manipulan y al dominio que estas comunidades tengan sobre un formato en específico. En este epígrafe se analizarán algunos de estos formatos, con el objetivo de conocer sus principales características y algunas de sus aplicaciones. Estos formatos son de

propósito general, por lo que pueden ser utilizados en múltiples áreas de aplicación diferentes de las tratadas en esta tesis (Muñiz Fernández *et al.*, 2012; Ullman y Denning, 2012; Castro *et al.*, 2013; Long *et al.*, 2013; Wang *et al.*, 2013).

De las herramientas presentadas en el epígrafe 2.6, ninguna de ellas está integrada con los formatos de datos científicos que se describen en los siguientes epígrafes.

2.7.1. HDF

Hierarchical Data Format, más conocido por HDF, fue desarrollado por el Centro Nacional de Aplicaciones de Supercómputo (*National Center for Supercomputing Applications, NCSA*) en el año 1988. En la actualidad, su soporte corre a cargo de HDF Group de la Universidad de Illinois. Es un formato de datos de propósito general, flexible, portable y eficiente para el almacenamiento y recuperación de datos científicos (Poinot, 2010; Ullman y Denning, 2012). HDF es también un conjunto de bibliotecas escritas en Java, eficiente en las operaciones de E/S debido al uso de paralelismo (Gray *et al.*, 2005); es libre, de código abierto y multiplataforma. Dado su excelente desempeño y simplicidad ha sido ampliamente usado por muchas comunidades de científicos (Long *et al.*, 2013). Este formato es muy usado por entidades que producen y gestionan información de carácter ambiental y otras entidades de observación territorial, como es el caso de la NASA (Vanegas, 2013).

HDF se encuentra disponible de forma libre. La distribución consiste en la biblioteca, utilidades de línea de comando, una suite de prueba, interfaces con Java y HDFView (Pfeiffer *et al.*, 2012), un visor basado en Java, mediante el cual los usuarios pueden observar fácilmente detalles interesantes de los datos. HDF presenta cuatro niveles de interacción. En su nivel más bajo es un archivo para el almacenamiento de datos científicos. En su nivel más alto es una colección de utilidades y aplicaciones para manipular, ver y analizar datos en los ficheros HDF, y en los niveles intermedios se encuentra una biblioteca de programas que provee APIs de alto nivel y una interfaz de datos de bajo nivel (obsérvese la figura 2.13)

Aplicaciones generales

En el nivel más alto hay utilidades de línea de comandos de HDF, aplicaciones de NCSA que soportan visualización de datos y análisis, y una variedad de aplicaciones de terceros desarrolladores. Existen utilidades de línea de comandos de HDF para:

- Convertir de un formato a otro (por ejemplo, desde y hacia JPEG/HDF)
- Analizar y ver ficheros HDF (siendo hdp, una de las herramientas más útiles)
- Manipular los ficheros HDF

De las utilidades de HDF, *hdp* es una de las más importantes, su función es proveer información rápida sobre los contenidos y objetos de datos en un archivo HDF (Long *et al.*, 2013),

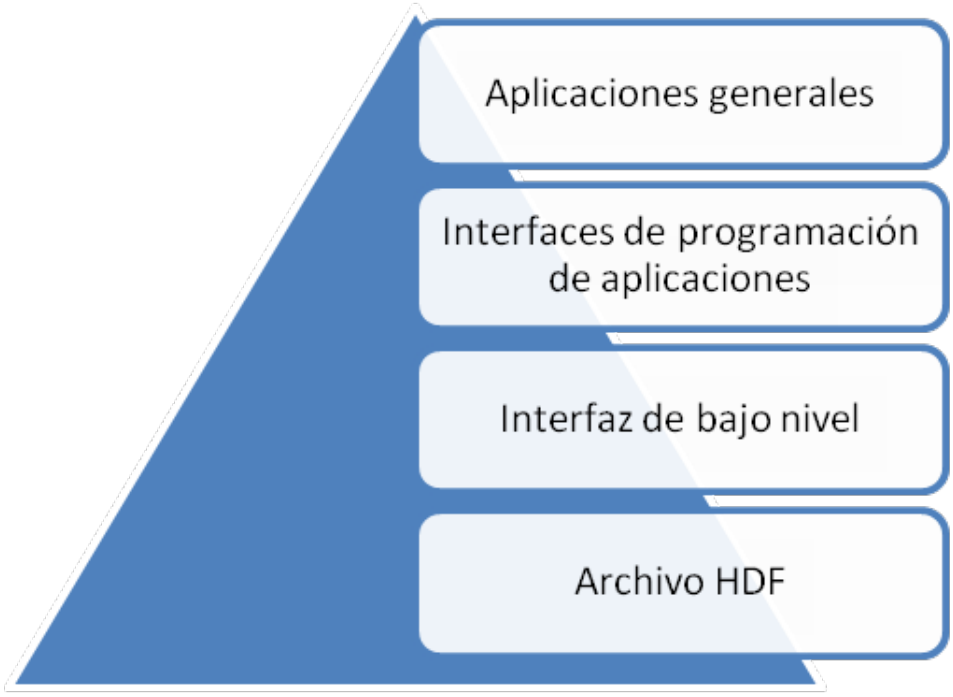

Figura 2.13 Niveles de interacción de HDF.

puede listar los contenidos de los archivos HDF en varios niveles con diferentes detalles, además puede también vaciar los datos de uno o más archivos, en formato binario o ASCII.

Compresión de datos

HDF4 y versiones posteriores como HDF5 (Wang *et al.*, 2013), soportan una interfaz de compresión de bajo nivel, la cual permite que cualquier objeto de dato sea comprimido utilizando una variedad de algoritmos. Actualmente están soportados solo tres algoritmos de compresión: *Run-length Encoding (RLE)*, Huffman adaptativo, y un codificador de diccionario *LZ-77* (el algoritmo de decodificación de gzip) (Yeh *et al.*, 2002). Los planes para algoritmos futuros incluyen un codificador de diccionario *Lempel/Ziv-78*, un codificador aritmético y un algoritmo rápido de Huffman. HDF4 y sus versiones posteriores soportan compresión de n-bit para *SDS, RLE* (*Run-Length Encoding*), IMCOMP, y compresión JPEG para imágenes de puntos.

El archivo HDF

HDF puede almacenar varios tipos de objetos de datos dentro de un archivo, como imágenes de mapa de bits, paletas, texto y estilos de tablas de datos. Cada *objeto* en un archivo HDF tiene una etiqueta predefinida que indica el tipo de dato y un número de referencia que identifica la instancia. Hay un número de etiquetas disponibles para la definición que hace el usuario sobre los tipos de datos. Sin embargo, solamente las personas que tienen acceso al **software** del usuario pueden acceder adecuadamente a los datos. Los usuarios de HDF no necesitan conocer el formato físico, pues el único método práctico de acceso y manipulación de los datos

es mediante interfaces de software. Existen diversas formas de acceder a los datos contenidos en un archivo HDF. Una de ellas es el acceso a nivel binario, si se conoce la estructura del archivo. A continuación se describe la estructura básica del formato HDF (Group, 2011):

- *File Header*: El encabezado es el primer componente de un archivo HDF que incluye los primeros cuatro bytes en el archivo HDF. El encabezado es una firma de los archivos HDF, con valores hexadecimales, por ejemplo, $0x0E, 0x03, 0x13, 0x01$.
- *RootGroup*: El grupo raíz es una analogía de lo que constituye un directorio de sistema de archivos. Un grupo tiene cero o más objetos y cada objeto debe ser miembro de al menos un grupo. El grupo raíz es un caso especial, pues no puede ser miembro de ningún grupo. Contiene los campos siguientes: nombre y atributos. Los atributos tienen asociada la información siguiente: nombre; valor; tipo, que pueden tomar los tipos de datos: *string, byte (8-bit), short (16-bit), int (32-bit), unsigned byte (8-bit), unsigned byte (16-bit), unsigned byte (32-bit), long (64-bit), float, double, objectreference*; tipo de almacenamiento: *Max string length* para tipos de datos *string* y *Arraysize* para todos los otros tipos de datos.
- *NamedObject*: Hay tres clases de nombres de objetos: Grupos, Conjuntos de datos y tipo de dato nombrado. Cada uno de esos objetos es miembro de al menos un Group y un enlace.
- *Group*: el grupo es similar al *RootGroup*.
- *Dataset*: Un conjunto de datos es un arreglo multidimensional (rectangular) de elementos de datos. El objeto del espacio de datos define la forma de la matriz (número de dimensiones y tamaño de cada dimensión). Se asocia con la siguiente información: nombre y grupo padre.
- *Datatype*: el tipo de dato describe las características de un único elemento de dato. Existen dos categorías de datos: atómicos y compuestos. Un tipo de dato atómico representa un objeto simple como: una cadena de caracteres, un carácter, un entero y números de punto flotante. Un tipo de dato compuesto se compone de múltiples elementos de tipos de datos atómicos, como: arreglo, enumerativo, tipos de tamaño de variable y complejo. Se asocia con la siguiente información: clase de tipo de datos, tamaño en *bits* y ordenamiento de *bytes*.
- *Dataspace*: El espacio de datos describe el diseño de los elementos de una matriz multidimensional. Describe el hiper-rectángulo como una lista de las dimensiones actuales y con los tamaños máximo (o ilimitado). Se asocia con la dimensión y tamaño de la matriz.
- *Attribute*: los atributos se utilizan para documentar los objetos. Los atributos de un objeto son el nombre y el dato. Un atributo es similar a un conjunto de datos (*Dataset*) pero con algunas limitaciones: Sólo se puede acceder por medio del objeto. El nombre de los atributos tiene significado sólo sin el objeto. Un atributo debe ser un objeto pequeño. Un acceso simple debe leer y escribir los datos de un atributo. Los atributos no tienen atributos.

HDF ha servido de base para la construcción de otros formatos de datos científicos, como es el caso de HDF-EOS, descrito en el siguiente epígrafe.

2.7.2. HDF-EOS

El sistema observatorio de la tierra (EOS por sus siglas en inglés de *Earth Observing System*) es un programa iniciado por la empresa de las ciencias de la tierra de la NASA para el estudio del cambio climático global. EOS usa principalmente instrumentos aerotransportados para obtener información sobre la tierra, la atmósfera, los océanos y los fenómenos físico-químicos que ocurren sobre el sistema terrestre (Yang y Di, 2004). Actualmente el programa EOS genera más de dos *terabytes* de datos por día (Yang y Di, 2004).

HDF-EOS es un formato estándar para los datos EOS producidos por la NASA. El formato HDF fue seleccionado como sistema base para almacenar los datos EOS; sin embargo, este carece de mecanismos para guardar información geo-localizada, es vital para los datos geoespaciales. El proyecto EOS amplió HDF a HDF-EOS para añadirle tres nuevos modelos de datos (*point, swath y grid*). De los tres modelos de datos añadidos, *swath* y *grid* pueden tener dos o múltiples dimensiones espaciales, mientras que *point* es un modelo de dato discreto. El modelo *grid* es usado para datos geo-rectificados, donde los elementos de datos espaciales (píxeles) son regularmente organizados y todos los píxeles tienen la misma forma y tamaño. Las coordenadas espaciales de cualquier pixel pueden ser obtenidas fácilmente, a partir de las coordenadas de un pixel de referencia, usualmente la esquina superior o inferior izquierda y el tamaño del pixel. Los datos *grid* pueden ser fácilmente añadidos a otros conjuntos de datos espaciales para ser analizados por sistemas de información geográfica y sensores remotos (Yang y Di, 2004).

El modelo de datos *swath* es para datos que van a ser geo-rectificados, donde diferentes píxeles tienen diferentes formas y tamaños. Sin embargo, sin ser geo-rectificados los datos swath no pueden ser combinados directamente con otros datos espaciales geo-rectificados. Actualmente muchos sistemas de información geográfica y sensores remotos no son capaces de procesar los HDF-EOS con modelos de datos *swath* debido a no poder geo-rectificar estos datos. Esto, sin lugar a duda, ha significado una limitante para muchos usuarios de EOS, por lo que actualmente se realizan estudios y trabajos con el objetivo de crear herramientas que ayuden a los usuarios a geo-rectificar sus datos.

2.7.3. CDF

Common Data Format (CDF) es un formato de datos para almacenar conjuntos de datos multidimensionales. Sus orígenes datan del desarrollo del Sistema de Datos Climáticos de la NASA en el *National Space Science Data Center (NSSDC)*. Según (NASA, 2013) una de las ca-

racterísticas más importantes de CDF es que puede manipular conjuntos de datos inherentemente dimensionales además de los escalares. Para lograr esto CDF agrupa los datos por variables cuyos valores son organizados conceptualmente en arreglos. Las variables CDF son nombres genéricos u objetos que representan datos. Por ejemplo, una variable puede representar datos de una variable independiente, una variable dependiente, el tiempo, un valor de fecha, etc. En otras palabras, las variables no contienen ningún significado aparte de los datos en sí mismos.

La dimensionalidad de una variable depende de cómo sean especificados los datos por los usuarios. Para un dato escalar, por ejemplo, el arreglo de valores debe ser *0-dimensional*; mientras que para los datos de una imagen sería de dos dimensiones. CDF permite a los usuarios especificar arreglos hasta de diez dimensiones. Los arreglos para una variable en particular son llamados registros variables. Una colección de arreglos, uno para cada variable se nombra como un registro CDF. Un CDF puede contener múltiples registros CDFs, los cuales son útiles para los datos que son observados en diferentes instantes de tiempo.

Mientras que las variables son un mecanismo para representar los datos, los atributos en un CDF son un mecanismo para describir el archivo CDF y sus variables. Existen dos tipos de atributos en un CDF: los atributos globales para describir el archivo CDF y los atributos de variables para describir cada variable. Ejemplos de atributos globales pueden ser: fecha de creación del archivo, autor, documentación de los datos, etc. Ejemplos de atributos de variables son: valor máximo y mínimo, unidades de medición de los datos, etc.

Es importante destacar que no existe solo una forma correcta de almacenar los datos en un CDF, dado que los usuarios tienen total control de cómo guardar los datos. Esta es una de las ventajas de CDF. Los datos pueden ser organizados de cualquier forma según crea el usuario. CDF es también una biblioteca flexible y extensible que brinda muchas facilidades para la creación y el acceso a un CDF (NASA, 2013); la misma permite crear los formatos de archivos para almacenar los datos y metadatos de dos maneras: la primera es el formato tradicional de multi-archivos CDF; de esta manera se crea un archivo *.cdf* con toda la información y metadatos, además de un archivo para cada variable *.v#*, donde # es el número de la variable. La segunda opción es crear un único archivo *.cdf* que contiene toda la información, metadatos y los valores de los datos para cada variable en el CDF. Ambos formatos permiten el acceso directo a los datos. La ventaja de un único archivo es que minimiza el número de archivos a manejar y facilita su distribución a través de redes. Esta biblioteca también permite la compresión de datos pero sólo para un CDF creado a partir de un único archivo.

2.7.4. NetCDF

El formato de archivo NetCDF (*Network Common Data Format*) fue implementado por el programa *Unidata*, uno de los ocho programas de la *University Corporation for Atmospheric Research (UCAR)*, y fue establecido como formato estándar de la comunidad científica para el almacenamiento de todo tipo de datos oceanográficos y atmosféricos desde el año 1989

(Borrell González, 2012). La fortaleza de NetCDF está dada por la sencillez de su modelo de datos subyacente, su flexibilidad y su eficiente acceso a los datos (Hankin *et al.*, 2010). El formato NetCDF contiene metadatos que ayudan a identificar qué clase de datos se encuentran almacenados; suelen contener información sobre el tipo de variable, unidades de medición y dimensiones. NetCDF, a diferencia de otros formatos, no requiere archivos adicionales para su interpretación.

Según UNIDATA (2012), los datos NetCDF tienen las características siguientes:

- Descripción automática de los datos mediante metadatos.
- Portabilidad y acceso a datos de diferentes formas.
- Eficiencia en el acceso escalable a pequeños subconjuntos de un gran conjunto de datos.
- Soporte para adicionar datos a estructuras previamente establecidas sin necesidad de copiar ni redefinir todo el conjunto de datos.
- Posibilidad de accesos simultáneos a datos por parte de múltiples procesos de escritura y lectura.
- Compatibilidad con versiones anteriores del formato.

Estándares que cumple NetCDF

NetCDF cumple con los estándares siguientes:

- OGC (*Open Geospatial Consortium*), persigue acuerdos entre las diferentes empresas del sector que posibiliten la interoperación de sus sistemas de geo-procesamiento y el intercambio de la información geográfica en beneficio de los usuarios.
- CDI (*Common Data Index*) tiene como objetivo brindar a los usuarios una visualización muy detallada y gran difusión geográfica de información oceanográfica a través de los datos obtenidos de diferentes instituciones.
- ISO 19115 (*International Organization for Standardization*), es la entidad internacional que se ocupa de promover el desarrollo de estándares internacionales para la construcción de productos, su comercialización y comunicación. Busca la estandarización de normas de productos para organizaciones a nivel internacional.

NetCDF tiene una interfaz de programación de aplicaciones (API) bien diseñada, que permite el desarrollo de aplicaciones en varios lenguajes de programación. Esta interfaz es usada también para una biblioteca de funciones de acceso a datos, que se almacenan y recuperan en forma de matrices (Rew *et al.*, 2011). A los valores de las matrices se puede acceder directamente, sin conocer detalles del almacenamiento de los datos. El desarrollo del formato NetCDF mejora la accesibilidad y la reutilización de los datos en las matrices.

Recientemente los desarrolladores de NetCDF y HDF han colaborado para crear un software que combina las mejores características de cada uno, lo que permite una mayor interoperabilidad entre las dos comunidades (Hankin *et al.*, 2010). El software resultante fue NetCDF-4,

compatible con versiones anteriores de programas y datos para NetCDF-3, el cual mejora el rendimiento e incrementa las capacidades de codificación de colecciones de datos.

El archivo NetCDF

Los archivos NetCDF tienen una estructura interna definida por el modelo común de datos (CDM *Common Data Model* por sus siglas en inglés). El CDM es un modelo de datos abstractos para conjuntos de datos científicos; este modelo fusiona los modelos NetCDF, OPeNDAP y HDF5 para crear una API común para muchos tipos de datos científicos.

El CDM consta de varias capas, donde cada una se va superponiendo en la parte superior de la anterior, para ir añadiendo cada vez una semántica más compleja:

- La capa de acceso de datos, se encarga de los datos de lectura y escritura.
- La capa de sistemas de coordenadas identifica las coordenadas de las matrices de datos.
- La capa de atributo estándar, es la que conoce algunos de los significados que los humanos utilizamos para hacer referencia a los datos científicos, tales como unidades, sistemas de coordenadas, topología de datos, etc.
- La capa de características de tipo de dato científico, identifica los tipos de datos con la adición de métodos especializados para cada tipo de datos.

Según Rew *et al.* (2011), los tipos de datos abstractos más importantes que componen un archivo NetCDF, lo constituyen:

- Conjunto de Datos (*Dataset*): Puede ser un NetCDF, HDF5, OPeNDAP o cualquier otro tipo de documento al que se pueda acceder a través de la API de NetCDF.
- Grupo: Un grupo contiene variables, dimensiones, atributos, etc. En definitiva, los grupos son el contenedor de todo aquello que contiene un conjunto de datos formando un árbol jerárquico. Por eso, como mínimo, siempre habrá un grupo definido en un archivo.
- ProtoVariable: En los archivos NetCDF una ProtoVariable es un contenedor de datos. Está compuesta por diversas variables que contienen su información, como por ejemplo el tipo de datos, su dimensión o dimensiones que definen su matriz y, opcionalmente, se pueden complementar añadiéndole un conjunto de atributos que la describan o definan aun con mayor exactitud.
- Dimensión: Una dimensión se utiliza para definir en función de qué van a variar los valores de nuestra variable. Puede ser compartida entre las demás variables, lo que proporciona una manera simple y eficaz de asociarlas.
- Atributo: Un atributo tiene un nombre y un valor que se utilizan para asociar metadatos con una variable o grupo. Por ejemplo, si se trata de un atributo de una variable estaríamos hablando de unidades o nombre estándar y solamente afectaría a esta variable. Si nos referimos a un atributo de un grupo, este estaría afectando a todo el conjunto de datos y el atributo sería, por ejemplo, para definir qué convención se ha utilizado o el nombre de la organización que ha generado estos datos.

- Estructura: Es un tipo de variable que contiene otras variables. De esta manera los datos almacenados en una estructura se encuentran físicamente muy juntos en el disco, lo que hace que sea eficaz la recuperación de estos datos al mismo tiempo.
- *Array*: Un arreglo (o matriz) contiene los datos reales de una variable obtenidos del disco, red, base de datos, etc. Podemos decir que un arreglo es el contenedor de las series de datos.

Otro de los formatos más utilizados para la manipulación de imágenes, se describe en el siguiente epígrafe, aunque actualmente no es muy utilizado para ser integrado con sistemas de información geográfica, en un futuro pudiera emplearse para realizar análisis espacio-temporales relacionados con la astronomía.

2.7.5. FITS

Flexible Image Transport System (FITS), es el formato de archivo más utilizado comúnmente en el mundo de la astronomía. A menudo se utiliza para almacenar y manipular datos que no corresponden precisamente con imágenes, como por ejemplo: listas de fotones, datos del espectro electromagnético, datos en tres dimensiones y muchos más. Un archivo FITS puede estar formado por varias ramificaciones, y cada una de ellas puede almacenar datos diferentes de un mismo objeto. Por ejemplo, se pueden almacenar varias bandas en un mismo archivo FITS, conteniendo mediciones en diferentes rangos de frecuencia del espacio electro magnético (Group, 2009).

Una de las mayores ventajas de FITS como formato de dato científico es que la información de las cabeceras de los archivos es legible en formato *ASCII*; de esta forma los usuarios pueden examinar las cabeceras para conocer lo que almacena un archivo determinado. Los archivos FITS, contienen una o más cabeceras, y cada una de ellas almacena secuencias de cadenas de caracteres fijos que llevan pares de valores. Estos son intercalados entre los bloques de datos y suministran metadatos como: tamaño y dimensiones de los datos, estructura interna, comentarios, historial de los procesos por los que han pasado estos datos desde su origen hasta su estructura actual y cualquier otra información de interés para los usuarios potenciales. Aunque FITS posee varias palabras restringidas, el estándar permite el uso arbitrario de todas las palabras. Está soportado mediante bibliotecas disponibles en los lenguajes más utilizados en el ámbito científico, incluyendo C, C++, C#, Fortran, Perl, Java, PDL, Pythone IDL (Hanisch *et al.*, 2000).

2.8. Conclusiones parciales

El estudio sistemático de los principales conceptos de la visualización científica y de las principales técnicas de visualización de datos multiparamétricos permitió seleccionar un grupo

de técnicas que son adecuadas para integrar en sistemas de información geográfica con el objetivo de realizar análisis exploratorio de datos. De las técnicas de visualización presentadas en este capítulo, se implementó un alto número de ellas, las cuales son tratadas en los capítulos 4 y 5.

Se estudiaron las principales características de algunos de los sistemas de información geográfica más populares, se valoraron características como la disponibilidad del código fuente para modificarlo con vistas a integrar las técnicas de visualización desarrolladas, la facilidad para incorporar nuevas funcionalidades y la documentación disponible para desarrolladores. De ellos se seleccionaron gvSIG y Sextante para probar los modelos propuestos en esta tesis. En el diseño e implementación de las herramientas se tuvieron en cuenta conceptos básicos del análisis exploratorio de datos; estos conceptos se valoraron desde la perspectiva de los datos, las tareas que se pueden llevar a cabo en el análisis exploratorio de datos, así como las principales herramientas que no deben faltar cuando se hace análisis exploratorio de datos. Los principios fundamentales de este tipo de análisis estuvieron muy presentes en cada una de las decisiones de diseño e implementación.

Luego de estudiar algunas de las principales herramientas que permiten realizar análisis exploratorio de datos mediante la integración de técnicas de visualización y el uso de mapas y formatos de sistemas de información geográfica, se pudieron analizar sus principales ventajas y deficiencias. Este estudio permitió realizar valoraciones sobre elementos novedosos en los que se debía trabajar. La manipulación de grandes volúmenes de datos científicos constituyó uno de los retos que emergieron del estudio de las herramientas existentes. Debido a esto se estudió además un conjunto de formatos de datos científicos que permiten la manipulación efectiva de un volumen considerable de información. De los formatos estudiados se seleccionaron HDF y netCDF para su integración con las herramientas de visualización en sistemas de información geográfica. Los principales algoritmos y herramientas que hacen uso de estos formatos de datos científicos en esta tesis son descritos en los capítulos 5 y 6.

Parte II

Contribuciones

3 Propuesta conceptual

En la última década se han obtenido grandes avances en el análisis visual de datos (Sacha *et al.*, 2014). Se realizaron numerosos estudios que demuestran el éxito de esta disciplina como ayuda a los expertos de dominio específico en la exploración de grandes y complejos conjuntos de datos. El poder del análisis visual de datos proviene de las habilidades perceptivas, el razonamiento cognitivo y el conocimiento del dominio en el lado de los humanos y la capacidad de cómputo y de almacenamiento de datos en el lado de los ordenadores, así como el perfecto acople que existe entre estos a través de las representaciones visuales.

En el análisis visual de datos, los datos se utilizan para arribar a conclusiones relacionadas con el mundo real, procesos, o un campo de aplicación. Anteriormente se mencionó la importancia de la participación de los humanos en este proceso, pues, por una parte, los ordenadores no poseen las cualidades creativas de los humanos que permiten encontrar conexiones interesantes, ocultas o sutiles entre los datos y el dominio del problema. Por otra parte, los humanos no son capaces de manipular grandes cantidades de datos de una manera eficiente y efectiva. Es por esto que en el análisis visual de datos se explotan las ventajas de ambos (humanos y ordenadores), para manipular grandes cantidades de datos, generar visualizaciones y aprovechar las habilidades en la percepción que poseen los humanos.

Respecto a las diferentes fases del diseño de herramientas de análisis visual, existen varios problemas sin resolver especialmente cuando se intentan acoplar en el análisis, datos espaciales y temporales. Esto constituye un campo que ha sido poco investigado.

Las deficiencias encontradas al analizar las herramientas de geovisualización estudiadas en el epígrafe 2.6 junto con la significativa carencia de modelos que permitan guiar a los desarrolladores de este tipo de herramientas, justifica el desarrollo de un modelo que posibilite integrar visualización científica en sistemas de información geográfica para el análisis exploratorio de datos espacio-temporales.

El modelo que se propone en esta tesis posibilita facilitar y guiar el desarrollo de herramientas de geovisualización, que permitan la extracción de conocimiento a partir de datos espacio-temporales de múltiples variables, tanto cuando se cuenta con datos con baja densidad espacial y amplios en el tiempo, como con datos con una alta densidad espacial amplios

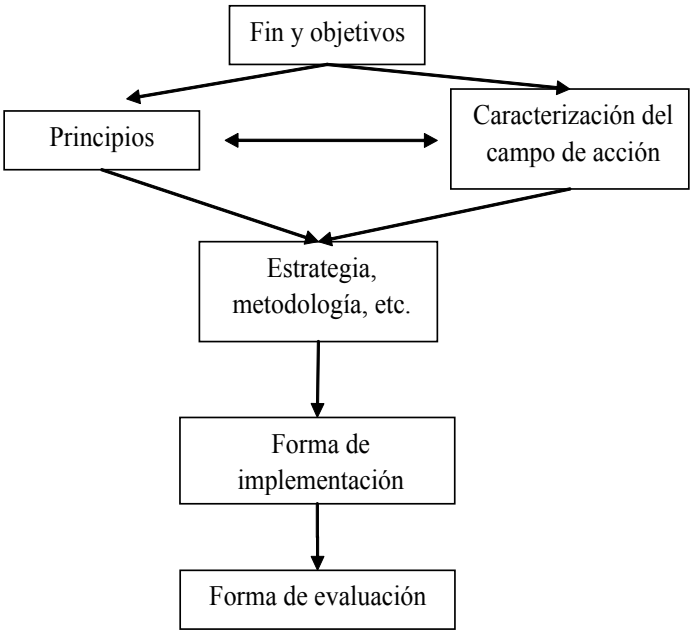

Figura 3.1 Componentes de un modelo según Valle-Lima (2012)

en el tiempo. Antes de presentar el modelo se hace necesario esclarecer algunos conceptos relacionados con el término modelo.

La palabra modelo proviene del latín *modŭlus*, que significa medida, ritmo, magnitud y está relacionada con la palabra *modus* que significa copia, imagen. El modelo y proceso que se sigue para llegar a este: la modelación se ha ido desarrollando y ampliando a tal punto, que en la actualidad se encuentran aplicaciones en disímiles esferas del saber (Valle-Lima, 2012).

Existen varias definiciones de modelo presentadas por múltiples autores, unos lo conciben en el plano mental (o sea, como abstracción) y otros en el plano material como reproducción a escala. En esta tesis se sigue la concepción de Valle-Lima (2012), en que ambas ideas pueden ser unidas. Para este investigador, un modelo puede asumir los componentes que se muestran en la figura 3.1:

- Fin y objetivos
- Principios
- Caracterización del objeto de investigación
- Estrategia (metodología, etc.)
- Formas de implementación
- Formas de evaluación

En este capítulo se definen cada uno de estos componentes con vistas a conformar el modelo propuesto. Del objeto de estudio se pueden inferir los componentes siguientes: el fin y los objetivos, los principios y la caracterización del deber ser del objeto de estudio propiamente. El fin y los objetivos establecen lo que se debe lograr con el modelo propuesto en relación con

el objeto de investigación sobre el cual se está trabajando. Los principios son regularidades más generales o esenciales que caracterizan el proceso o fenómeno en estudio y que guían la dirección de la transformación de este.

La caracterización del objeto de investigación (objeto, fenómeno o proceso que se investiga) es esencial para poder trabajar después en la evaluación del modelo. La estrategia se conforma por etapas, analizando para cada una de estas: su objetivo, una caracterización y las acciones concretas que se deben desarrollar; va dirigida hacia la transformación del objeto de estudio. Las formas de implementación son aquellas acciones que tienen como fin poner en práctica el modelo que se propone. Las formas de evaluación son las acciones que tienen como fin esencial analizar para emitir juicios de valor sobre el desarrollo de la aplicación y los resultados de la estrategia (Valle-Lima, 2012).

Se asume que el modelo es un reflejo mediatizado de la realidad sobre la base del cual opera el científico en ausencia del objeto para estudiarlo y explicarlo. Considera que el modelo revela una determinada unidad entre lo objetivo y lo subjetivo. Permite operar de manera práctica con un objeto o fenómeno, no de manera directa, sino utilizando el modelo como sustituto del objeto.

3.1. Esquema conceptual y modelo propuesto

A partir del objeto de estudio de esta investigación, que es el análisis visual de grandes volúmenes de datos espacio-temporales, se puede definir que el fin y los objetivos del modelo propuesto es facilitar y guiar el desarrollo de herramientas de *software* para el análisis visual de grandes volúmenes de datos espacio-temporales, mediante la integración de técnicas de visualización científica y sistemas de información geográfica. En este punto se trabaja en los aspectos siguientes:

- Análisis de visualizaciones y búsqueda de conglomerados. En particular, la visualización interactiva de las propiedades de conglomerados.
- Tratamiento y análisis de datos multivariados. Apoyo visual en el estudio de fenómenos espacio-temporales.
- Manipulación visual de grandes volúmenes de datos mediante técnicas de visualización científica a través de formatos de datos científicos integrados en sistemas de información geográfica.
- Desarrollo de funcionalidades que faciliten el proceso de generación de imágenes, facilitando la consulta, el filtrado y la interpretación de los datos.
- Concepción de representaciones más informativas y enfocadas en el objeto de análisis.

Como consecuencia, la primera prioridad está en el diseño de herramientas de nuevo tipo que permitan realizar análisis mediante visualizaciones más intuitivas, e interactivas, y brinden

la posibilidad de manipular grandes volúmenes de datos y gestionar múltiples vistas de estos de forma simultánea, teniendo en cuenta sus características espacio-temporales. El apoyo al usuario en la especificación de problemas analíticos y la determinación de las técnicas de visualización adecuadas proporcionará un nivel superior de análisis en este objeto de estudio.

Los principios o generalidades esenciales que caracterizan el análisis visual de grandes volúmenes de datos espacio-temporales están estrechamente relacionados con los principios presentados en el epígrafe 2.5.4. Se han tomado como base estos principios y se han adaptado al contexto del modelo:

- Ver el todo. En este caso, ver el todo se refiere a todos los objetos referencia, según la idea presentada en el epígrafe 2.5.1. En el contexto espacio-temporal las referencias pueden ser el espacio y el tiempo, por lo que se debe permitir representar los datos de tal manera que se pueda percibir el comportamiento general por medio de la visión de los momentos de tiempo y lugares del espacio que puedan estar involucrados en un análisis.
- Simplificar y resumir, así como dividir y agrupar. Estos son principios que tienen que ver con la granularidad, es decir, se debe brindar la posibilidad de manipular diferentes niveles de agrupación en los datos, sumarizaciones, simplificaciones, etc. En el contexto temporal esto implica poder analizar diferentes intervalos de tiempo, por ejemplo a nivel de minutos, horas, días, meses, años, etc., y concentrar el análisis en cada uno de estos niveles. En el contexto espacial la granularidad se puede ver, por ejemplo, como el paso de un análisis a nivel de municipios, provincias, regiones, etc., con que se quiera trabajar.
- Ver relaciones y elementos reconocibles. En el contexto geoespacial la posibilidad de ver relaciones y elementos reconocibles se brinda en los sistemas de información geográfica mediante la visualización de mapas (vectoriales y *raster*), la utilización de simbología y generación de gráficos simples. Mediante herramientas de visualización científica se pueden observar semejanzas, diferencias, comportamientos o fenómenos esperados entre diferentes atributos o grupos de atributos. Para esto se necesita un conocimiento previo de la estructura de los datos. Estos principios en ambos contextos deben estar integrados para realizar análisis espacio-temporales.
- Ampliar, enfocar y atender a particulares. En el contexto geoespacial, los sistemas de información geográfica permiten realizar operaciones de ampliar, enfocar y consultar información puntual sobre los mapas. Por otra parte, las herramientas de visualización también brindan este tipo de opciones: permiten a los analistas profundizar en el estudio de los diferentes gráficos y obtener referencias a datos puntuales que se destaquen. En el desarrollo de herramientas espacio-temporales también se debe respetar estos principios.
- Establecer enlaces. En sistemas de información geográfica este principio se manifiesta mediante el enlace de múltiples vistas relacionadas con los datos (por ejemplo, tablas, mapas, gráficos simples, leyendas, etc.). En la visualización científica se enlazan vistas de representaciones visuales de los datos bajo el concepto de múltiples vistas enlazadas (*Multiple*

Linked Views, por sus siglas en inglés). Este principio, en el contexto espacio-temporal, debe permitir manipular ambas opciones.

- Establecer estructura. Este principio se evidencia en los casos en que se sospecha que el comportamiento general puede estar asociado a ciertos patrones conocidos, y se distribuye y organiza de forma específica; o en los casos en que la interacción entre las referencias tiene conducta estructural, como pueden ser, procesos lineales o cíclicos en el tiempo, geolocalización de las referencias relacionadas con elementos espaciales; en cualquiera de ellos, debe permitirse explorar cada componente y sus interacciones con otros mediante la división de referencias relevantes en varias referencias, o introduciendo referencias adicionales.
- Involucrar conocimiento sobre el dominio. Este principio está más orientado hacia los analistas o usuarios de herramientas de *software*. Desde el punto de vista de los desarrolladores de herramientas, existe un grupo de acciones que pudieran tributar a involucrar conocimiento sobre el dominio. Por ejemplo, posibilitar la interacción con bases de datos de hechos y ontologías, el diseño y propuesta de posibles patrones y parámetros para las visualizaciones, así como el etiquetado de datos y la generación de metadatos.

El campo de acción es el contexto de donde surge la problemática. Para caracterizarlo, es necesario retomar la problemática que originó el objeto de estudio de esta investigación. Esta en general está dada por las dificultades que poseen los humanos para analizar numéricamente datos científicos relacionados con el espacio y el tiempo. Especialmente cuando se cuenta con grandes volúmenes de datos y múltiples variables, los cuales pueden contener series temporales. En la caracterización del campo de acción se encuentran los elementos siguientes:

- Las técnicas de visualización científica. En particular, las técnicas de visualización de datos multiparamétricos son de gran utilidad para abordar los problemas planteados. En el epígrafe 2.3 se realizó una descripción de las principales técnicas de visualización, que pueden ser utilizadas para visualizar múltiples variables simultáneamente. Se recomienda hacer un estudio para seleccionar teniendo en cuenta un problema dado, cuáles de las presentadas o reportadas en la literatura se ajustan más al problema. Estas técnicas pueden reunirse en bibliotecas de técnicas de visualización, modificarse y ocasionalmente se pueden crear nuevas técnicas para resolver diferentes problemas.
- Los sistemas de información geográfica. Particularmente los basados en *software* libre constituyen poderosas herramientas para el tratamiento, manipulación y visualización de datos espaciales. La ventaja de disponer de su código fuente permite que puedan ser extendidos con nuevas funcionalidades. En el epígrafe 2.4 se describieron algunos de los sistemas de información geográfica más populares en la actualidad. Se recomienda seleccionar sistemas de información geográfica que sean basados en software libre y que posean facilidades para la incorporación de nuevos módulos y la integración con otras

bibliotecas. La compatibilidad en cuanto a la integración es un factor importante a tener en cuenta. Algunos factores como el uso de licencias, el lenguaje de programación, la calidad de la documentación para desarrolladores y la forma de manipular los datos, pueden influir en la selección de un sistema de información geográfica para proceder a su modificación.

- Las tareas propias del análisis visual. Las preguntas que se deben responder mediante el análisis de datos constituyen un factor importante que debe tenerse en cuenta en el análisis visual de grandes volúmenes de datos. Estas fueron tratadas en el epígrafe 2.5.2. Se debe brindar la posibilidad de poder realizar diferentes tipos de tareas tanto simples como sinópticas. La forma de realizar las tareas debe ser intuitiva. Además, deben existir mecanismos para corroborar hipótesis mediante la proposición automática, o semi-automática de secuencias de tareas.
- Herramientas de *software* para el análisis visual de datos. En el epígrafe 2.5.2 se describieron 5 tipos de herramientas generales que se deben tener en cuenta para el análisis visual de datos. Algunas se enfocan más en la visualización y otras se enfocan más en los datos. Las interrelaciones que se pueden hacer entre ellas es muy variable y dependen del problema y del objetivo de cada análisis.
- Formatos de datos científicos. Un factor esencial para la manipulación de grandes volúmenes de datos espacio-temporales lo constituyen los formatos de datos científicos. En el epígrafe 2.7 se presentaron las principales bondades y características de algunos de los más populares. Estos formatos son más predominantes en las herramientas de visualización científica que en los sistemas de información geográfica, pero en los últimos años se ha notado un incremento en la integración de ellos con sistemas de información geográfica. Se propone la utilización de bibliotecas de estos formatos para facilitar la integración de datos, así como el almacenamiento y procesamiento.
- Actores. Se distinguen dos tipos de actores en este campo de acción. Por una parte se encuentran los desarrolladores de herramientas, y por otra, los analistas. Los desarrolladores de herramientas son los encargados de diseñar e implementar herramientas de geovisualización para el objeto de estudio. Los analistas, son los usuarios potenciales de estas herramientas. Ambos actores deben tener en cuenta los principios generales descritos en esta sección.

Estos son los principales elementos con que se aborda en esta investigación el análisis visual de grandes volúmenes de datos. Una vez definida la caracterización del campo de acción, se puede definir el resto de los elementos del modelo.

La estrategia o metodología está constituida por las vías o las formas que se deben seguir para llegar al objetivo. Anteriormente se mencionó que se puede dividir en etapas, donde cada una tiene su propio objetivo, una caracterización y un grupo de acciones que deben realizarse.

Se definen dos vías generales para obtener herramientas para el análisis visual de grandes volúmenes de datos espacio-temporales, especialmente si se tiene en cuenta que no se desea

desarrollar una herramienta de principio a fin, sino utilizar componentes de *software* ya desarrollados. La primera es incorporar en sistemas de información geográfica herramientas de visualización. La segunda es, a partir de sistemas de visualización, incorporarle funcionalidades para el soporte de información geográfica. Ambas variantes se pueden dividir en etapas más concretas que guíen y establezcan una secuencia para alcanzar un resultado.

Para la primera vía (obsérvese la figura 3.2) se debe asumir que se dispone de un conjunto de sistemas de información geográfica extensibles. La primera etapa para seguir esta vía se le puede llamar "selección del núcleo SIG". Esta etapa tiene como objetivo seleccionar un sistema de información geográfica adecuado para comenzar el estudio del problema. Las principales acciones en esta etapa están encaminadas a analizar los formatos de datos propios de los sistemas de información geográfica disponibles (*raster*, vectorial, tabular). Las acciones incluyen: el estudio de la compatibilidad de estos formatos con los datos del problema; el análisis de la documentación para desarrolladores de estos sistemas de información geográfica; así como la comprensión de los códigos fuentes con vistas a analizar las facilidades de modificación que lleven a la selección del sistema de información geográfica. La segunda etapa, "estudio de las carencias", tiene como objetivo delimitar los componentes que le faltan al sistema de información geográfica seleccionado para solucionar el problema. Las acciones de esta etapa están encaminadas a estudiar y seleccionar las técnicas de visualización más adecuadas para la extracción de conocimiento. Las acciones incluyen la selección de las tareas propias del análisis visual que deben estar involucradas en el sistema, así como las herramientas de *software* que se deban involucrar en el análisis visual de datos. En caso de que el sistema de información geográfica seleccionado no tenga soporte para la manipulación de formatos de datos científicos, se puede realizar un estudio para llevar a cabo la selección de los formatos que se puedan incorporar con mayor facilidad.

El objetivo de la tercera etapa es la integración de todas las tecnologías. Las acciones en esta etapa pueden ser: integración de bibliotecas en el sistema de información geográfica, desarrollo de nuevas funcionalidades para llevar a cabo tareas, y facilitar el uso de herramientas que cumplan con los principios propuestos.

La segunda vía sigue un esquema similar, en esta se asume que se dispone de un conjunto de sistemas de visualización científica, algunos de los cuales son de propósito general y permiten el análisis con una gran cantidad de técnicas de visualización. En ocasiones estos sistemas transforman sus datos a sus formatos propios y poseen herramientas para importarlos desde formatos de datos científicos.

La primera etapa de esta vía (obsérvese la figura 3.3) tiene como objetivo seleccionar un núcleo de un sistema de visualización que sea extensible, preferiblemente libre y de código abierto. Las principales tareas para esta etapa están encaminadas a estudiar las características de los sistemas de visualización, la compatibilidad con los datos del problema, soportar transformaciones entre formatos de datos geográficos y científicos. Las acciones incluyen además el estudio de la documentación para desarrolladores y las facilidades de extensión. La etapa

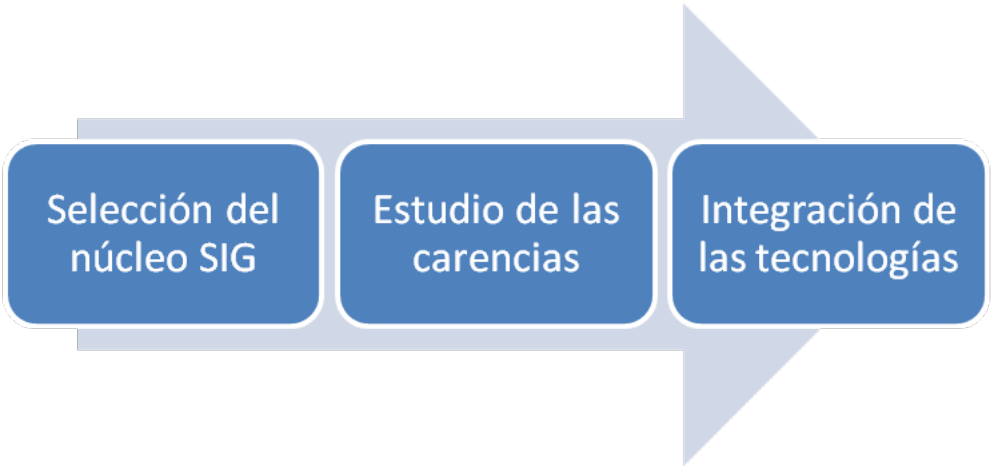

Figura 3.2 Primera vía para la construcción de herramientas que permitan el análisis visual de grandes volúmenes de datos espacio-temporales.

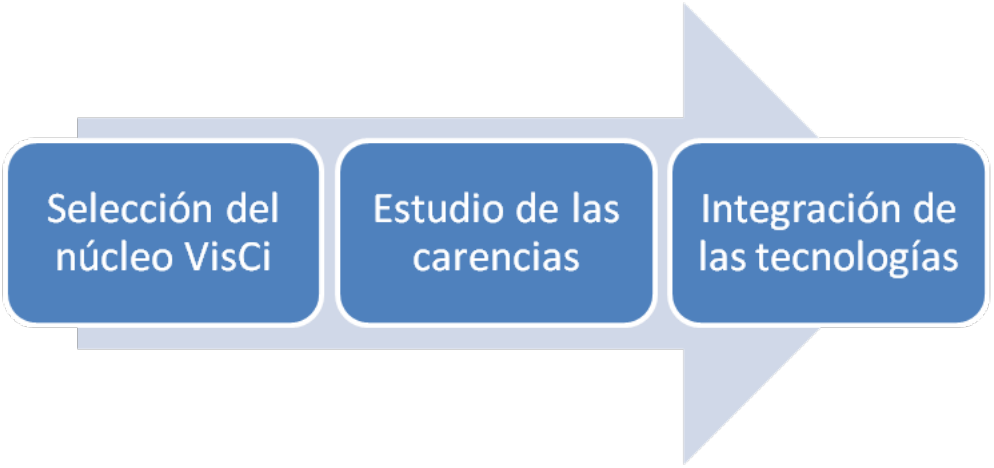

Figura 3.3 Segunda vía para la construcción de herramientas que permitan el análisis visual de grandes volúmenes de datos espacio-temporales.

concluye con la selección de un sistema de visualización que incorpore un grupo de técnicas que soporten análisis espaciales o temporales.

La segunda etapa, estudio de las carencias, tiene como objetivo descubrir los componentes que le faltan al sistema de visualización para soportar formatos de datos espaciales y temporales. Las acciones en esta etapa están encaminadas hacia el uso y desarrollo de módulos para soportar formatos de datos geográficos y tabulares, adecuar las tareas y las herramientas del sistema de visualización para soportar el análisis de datos espacio-temporales, teniendo en cuenta los principios generales del modelo propuesto. El objetivo de la tercera etapa es la integración de todas las tecnologías. Las acciones en esta etapa pueden ser, la integración de módulos para la importación y exportación de formatos de datos geográficos y tabulares en el sistema de visualización, el desarrollo de nuevas funcionalidades para llevar a cabo tareas, y facilitar el uso de herramientas que cumplan con los principios propuestos.

Otros de los componentes del modelo que se propone en esta tesis, es la forma de implementación. Anteriormente se mencionó, que son las acciones que tienen como fin poner en

práctica el modelo. Entre las principales acciones se encuentran: el estudio de las condicionantes y el contexto de un problema dado; el estudio y la representación de los datos; la definición de una arquitectura general para solucionar el problema por alguna de las vías de la metodología propuesta; la selección de los modelos de datos y métodos para llevar a cabo la solución de un problema; llevar a cabo casos de estudio para validar la solución del problema. En los capítulos 4 y 5 se muestra la ejecución del modelo para solucionar dos problemas diferentes. Las herramientas y algoritmos presentados en el capítulo 6, facilitan esta forma de implementación.

Las formas de evaluación son las acciones que tienen como fin esencial realizar análisis para emitir juicios de valor sobre el desarrollo de la aplicación y los resultados de la estrategia. En la sección 7.2 realiza una validación de los resultados mediante el análisis de encuestas que se presentaron a especialistas que interactuaron con las herramientas y métodos propuestos en este modelo. La opinión de estos especialistas, de conjunto con los resultados obtenidos en los casos de estudio mostrados en los capítulos 4, 5, 6 y 7, puede ser considerada como una forma de evaluación del modelo.

En la figura 3.4 se presenta el esquema conceptual del modelo propuesto donde se muestra cada uno de los componentes explicados anteriormente y sus relaciones. Se han utilizado las siglas HAVGVDET, para referirse a Herramientas para Análisis Visual de Grandes Volúmenes de Datos Espacio-Temporales.

3.2. Aplicación del modelo

Para la aplicación de este modelo es recomendable la definición de algunos componentes estructurales. Estos componentes se describen en el resto de los epígrafes de este capítulo. En la figura 3.5 se presentan los componentes básicos del funcionamiento de un sistema para el análisis visual de datos. Ese gráfico en el modelo propuesto está estrechamente relacionado con las técnicas de visualización científica; elemento esencial que se presenta en el recuadro caracterización del campo de acción del modelo.

El usuario o analista interactúa con la interfaz gráfica de usuario (GUI por sus siglas en inglés). Tiene la posibilidad de realizar la gestión del objetivo de un análisis y la gestión de metadatos, como muestran los recuadros correspondientes. En este caso el usuario puede ser considerado uno de los actores descritos en el modelo.

En la gestión del objetivo del análisis entra en juego el diseño de las técnicas de visualización de datos multiparamétricos (TVDM); este componente se encarga de la selección de las técnicas y de la parametrización correspondiente. Esto se logra haciendo uso de las bibliotecas de TVDM disponibles, las cuales incluyen las técnicas de visualización y los procedimientos estadísticos. Todos estos componentes se pueden ver como algoritmos básicos de procesamiento. En el diseño de estas técnicas se tienen que tener en cuenta los principios del modelo. En el diseño de estas técnicas entra en juego otro tipo de actor, los desarrolladores de herramientas. Estos,

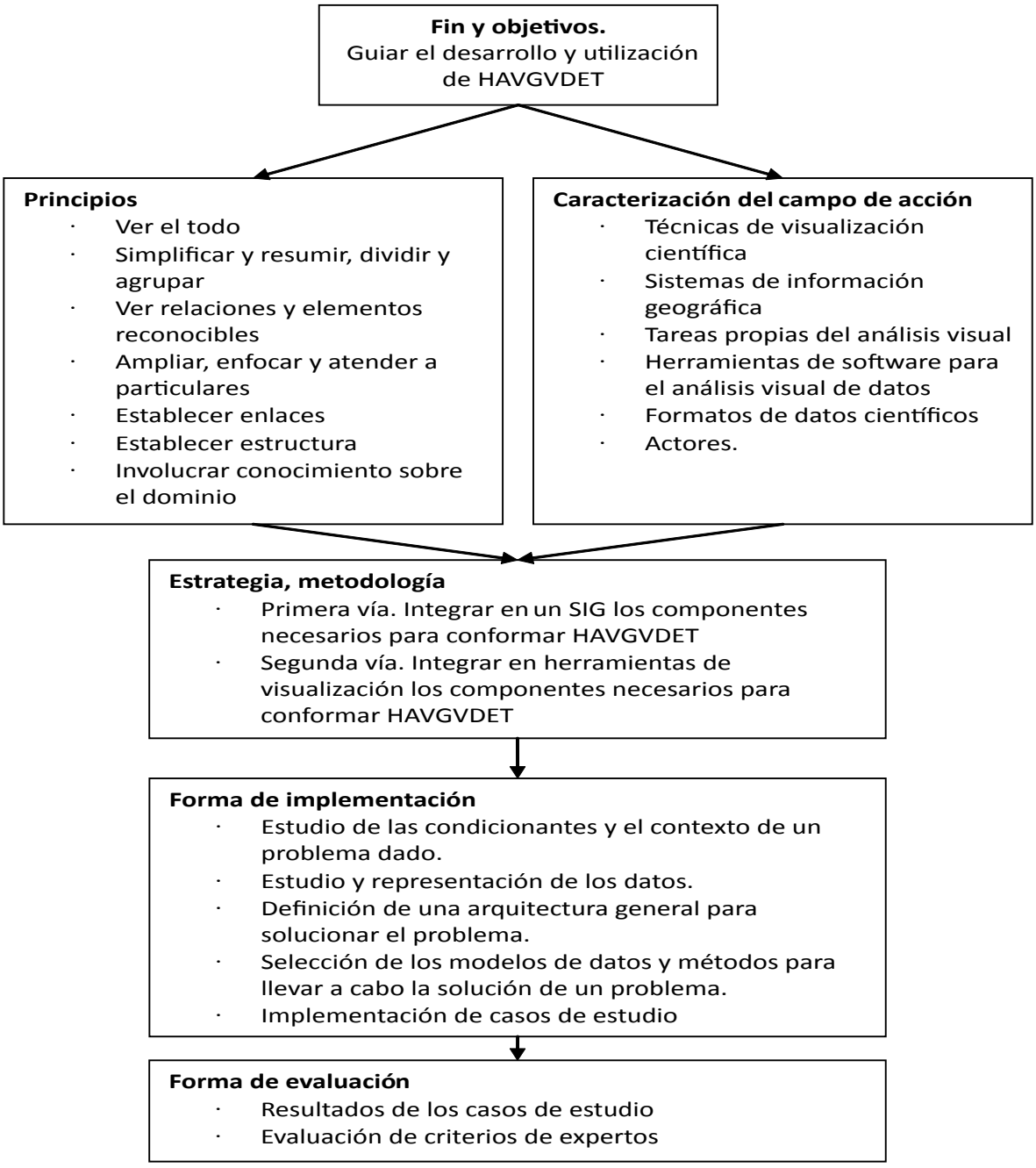

Figura 3.4 Esquema conceptual del modelo propuesto.

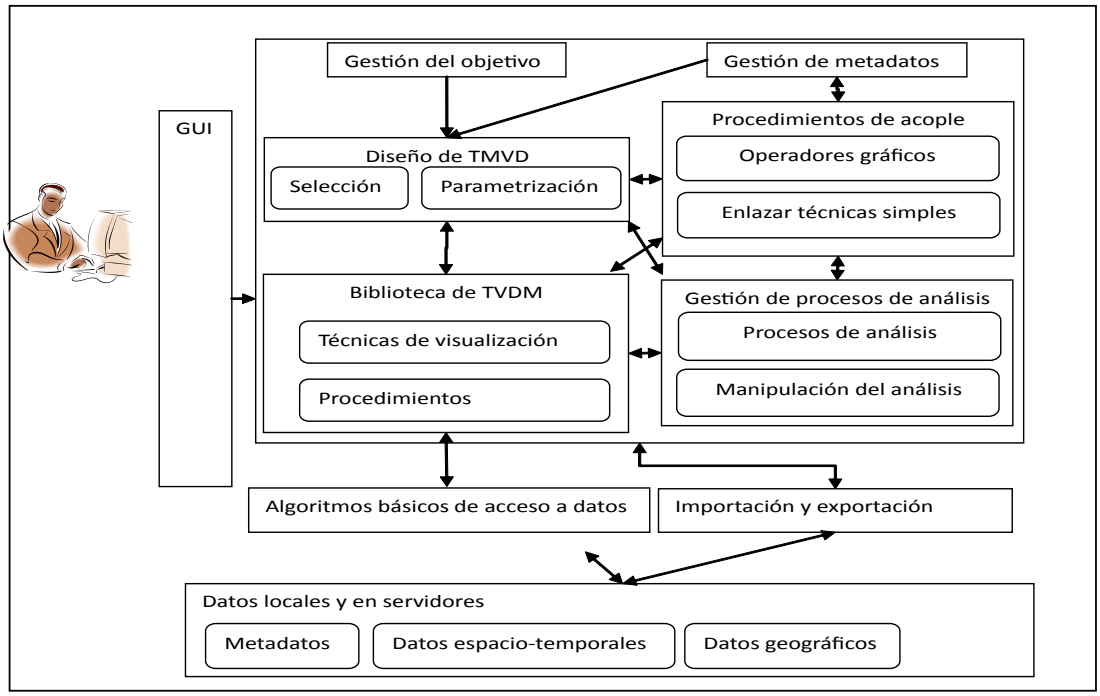

Figura 3.5 Componentes básicos del funcionamiento de un sistema para el análisis visual de datos.

basados en los principios tienen que ser capaces de integrar algunos de los componentes de la caracterización del campo de acción para dar solución a un problema concreto.

Por otra parte, la gestión de metadatos involucra, en primer lugar, los procedimientos de acople. Es decir, la forma en que se pueden enlazar técnicas de visualización simples y cómo orientar la utilización de los operadores gráficos. La gestión de procesos de análisis es otro proceso que involucra a la gestión de metadatos, esto incluye a los procesos de análisis en sí y la manipulación del análisis. En su base se encuentra la importación y manipulación de los datos. Los procedimientos de acoples están relacionados con varios componentes de modelo propuesto, las técnicas de visualización científica, los sistemas de información geográfica y las herramientas de *software* para el análisis visual. La relación de todos estos componentes de la caracterización del campo de acción facilita que exista una integración de datos y metadatos.

En la parte inferior del esquema se encuentran los datos, los cuales están en servidores o en máquinas locales. Dentro de este recuadro se encuentran los metadatos, datos espacio-temporales, datos geográficos y reglas. Todos están relacionados con el componente importación y exportación. Los componentes del modelo involucrados directamente en este aspecto son los componentes de *software* para el análisis visual de datos y los formatos de datos científicos, especialmente cuando se manipulan grandes volúmenes de datos, los sistemas de información geográfica suelen tener herramientas para la importación y exportación de datos a sus formatos propios.

En el modelo propuesto se propone un nuevo enfoque para la comparación de visualizaciones. La figura 3.6 muestra los principales componentes que se tienen en cuenta para este enfoque.

A partir de dos conjuntos de datos diferentes, se procede con el filtrado, selección y pro-

yección de los mismos, produciéndose a nivel de imagen una visualización de cada conjunto de datos. Los componentes del modelo que se involucran para este caso son las técnicas de visualización científica y las herramientas de *software* para el análisis visual de datos. Luego es que se tienen en cuenta las tareas propias del análisis visual, interactuando con las imágenes. Esta primera parte la explican los recuadros de la parte superior de la imagen.

El recuadro que tiene que ver con el nivel de los datos, contiene dos componentes principales, el primero tiene que ver con datos en forma de tablas y la manipulación de los formatos geográficos, tanto en formato vectorial como en formato *raster*, y el segundo con la visualización de mapas. Es importante tener en cuenta, una referencia espacial común, para que en los análisis exista una correcta transformación de coordenadas, de forma tal que se puedan homogeneizar las condiciones para realizar las comparaciones. Aquí intervienen los sistemas de información geográfica, las herramientas de *software* para el análisis visual de datos y los formatos de datos científicos.

Sobre los datos se pueden definir y seleccionar regiones que sean de interés para un usuario que busca dónde enfocar su análisis. De esta forma se pueden efectuar las visualizaciones mediante la combinación de mapas y técnicas de visualización. Las herramientas propias del análisis visual de datos se pueden tener en cuenta para este paso, así como los diferentes procedimientos de acople que se mencionaron en los componentes básicos del análisis visual de datos.

Un aspecto que no debe faltar es la leyenda, donde se deben mostrar componentes de visualización, interacción y algoritmos semiautomáticos que guíen a los usuarios en un análisis eficiente.

El diagrama general de flujo que deben seguir los datos tiene la estructura que se muestra en la figura 3.7. Se parte por la lectura de los datos, aquí entran en acción los procesos de importación y exportación. Luego se seleccionan las variables y atributos generales que se tendrán en cuenta para un análisis. En este paso intervienen algunas de las herramientas de *software* para el análisis de datos como las consultas y herramientas de filtrado.

En el siguiente paso se selecciona la cantidad de información necesaria seguido de la selección del espacio de observación. Los componentes del modelo involucrados en este caso son los sistemas de información geográfica y los formatos de datos científicos, mediante el uso de ellos se deben filtrar los datos para seleccionar variables, cantidades de registros y regiones de interés. Se debe definir cómo se hará la interpretación de la clase de datos. Es decir, las tareas propias del análisis visual que se tendrán en cuenta para los análisis. El siguiente paso corresponde con la definición del flujo de datos, donde intervienen las herramientas para la definición de flujos de trabajo. Por último se almacenan los datos en archivos, donde pueden intervenir los formatos de datos científicos.

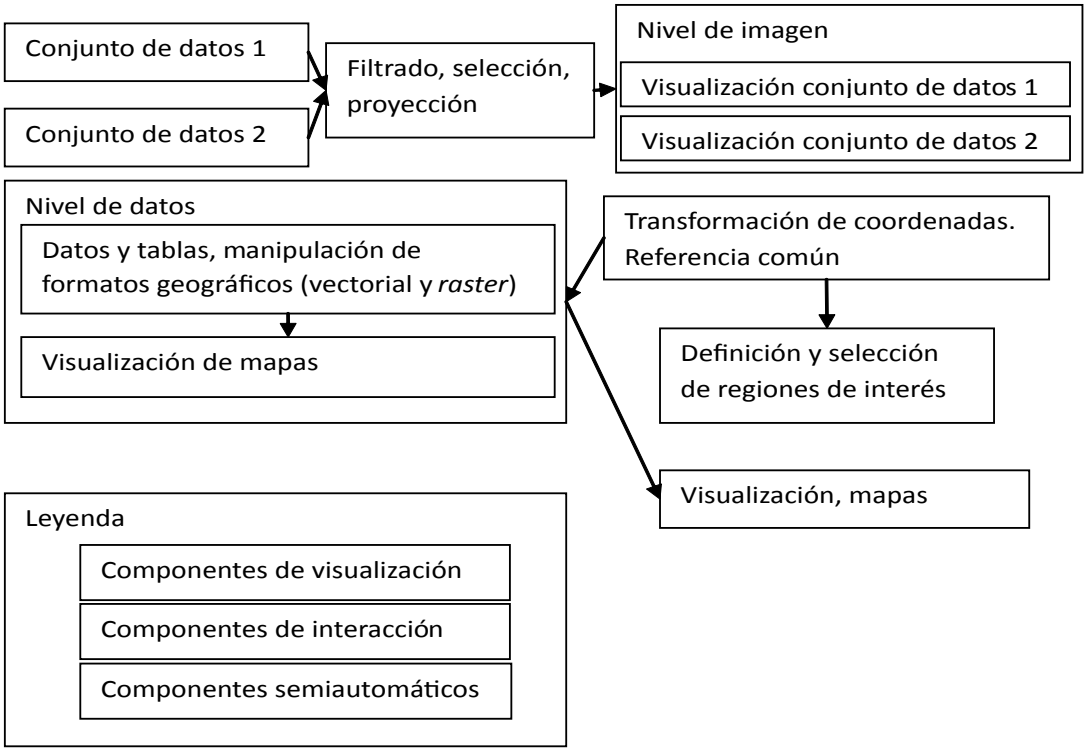

Figura 3.6 Nuevo enfoque para la comparación de visualizaciones.

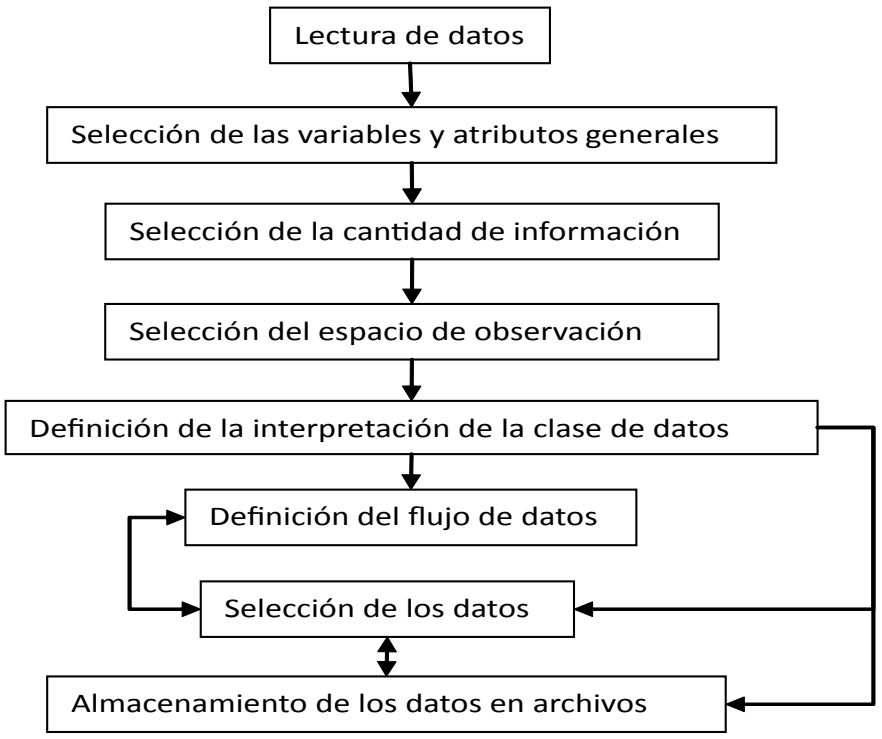

Figura 3.7 Diagrama de flujo general de los datos.

3.3. Visualización dinámica de datos

En esta tesis la visualización dinámica de datos se estudia a partir de tres enfoques: un primer enfoque está relacionado con la visualización de conjuntos de datos independientes, que no tienen por qué tener una referencia espacial. Corresponden con el diagrama de flujo de la parte superior de la figura 3.8. En este caso, la idea es utilizar un conjunto de datos multiparamétricos y realizar una visualización del conjunto de datos donde no se tiene conocimiento sobre la referencia espacial del conjunto. La relación de este primer enfoque con el modelo propuesto solo tiene que ver con las técnicas de visualización científica, los principios, las tareas propias del análisis visual de datos y las herramientas de *software* para el análisis visual de datos.

El diagrama de flujo del centro de la figura 3.8 corresponde con el segundo caso. Se trata de visualizar varios conjuntos de datos a la vez, obteniendo una percepción visual del origen de los datos, es decir, realizar las visualizaciones sobre un número pequeño de conjuntos de datos, plasmándolas sobre un mapa o manteniendo una referencia al lugar que corresponde con los datos. En este caso se adiciona el uso de herramientas de sistemas de información geográfica. Estos dos primeros casos son tratados y explicados en el capítulo 4, donde se aborda el análisis exploratorio de datos con baja densidad espacial. Aquí se ha seguido una de las vías de la estrategia propuesta en el modelo y se ha creado un conjunto de herramientas para solucionar un problema.

En la parte baja de la figura 3.8 se muestra un diagrama de flujo para el caso de la visualización de secuencias de datos almacenados en forma de mallas regulares, el cual contempla el análisis exploratorio de datos con alta densidad espacial. En cada celda del espacio geográfico se cuenta con un conjunto de datos multiparamétricos que puede almacenar datos temporales. En la relación con el modelo se incluyen los elementos anteriores más el uso de formatos de datos científicos. En el capítulo 5 se presenta la ejecución de una de las vías de la estrategia para la solución de otro problema diferente del presentado en el capítulo 4.

Como parte del seguimiento de una de las vías propuestas en el modelo, se construyó un *framework* general para el análisis exploratorio de grandes volúmenes de datos. La figura 3.9 muestra el Framework extScientificVisualization para gvSIG. La interfaz gráfica de usuario contiene un diálogo de parámetros generales, donde se configuran elementos para todas las técnicas de visualización, por ejemplo, la selección de los atributos que se desea involucrar en el análisis, un filtro determinado, etc. El modelizador gráfico se puede utilizar para realizar transformaciones de los datos en formatos de datos científicos, crear conjuntos de datos con la estructura necesaria para que pueda ser utilizado por el módulo de visualización. La vista de las configuraciones de las técnicas de visualización es donde se especifican las técnicas que van a ser involucradas en un análisis, sobre el mapa o en paneles independientes. En la vista de visualización se muestran todas las visualizaciones seleccionadas para un análisis.

Desde la perspectiva de los datos estos se tienen que manipular, ya sea desde el diálogo de

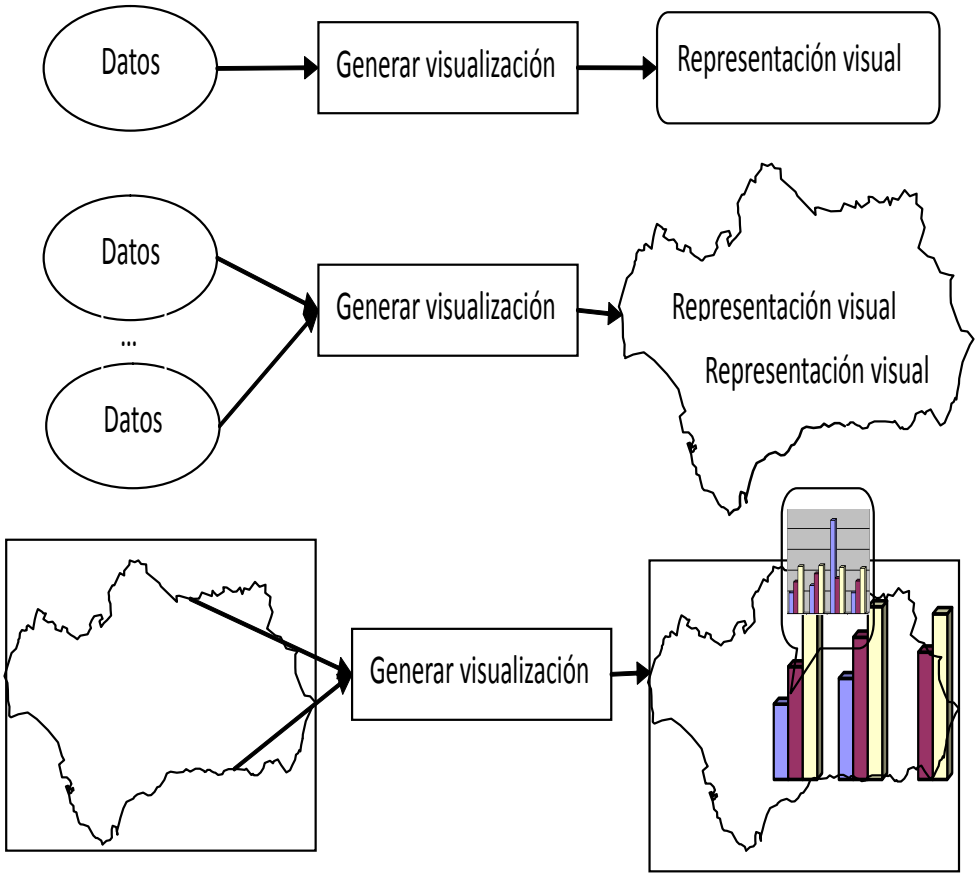

Figura 3.8 Casos para la visualización exploratoria de datos.

parámetros generales o desde los algoritmos de transformación de datos. La configuración de los parámetros de las visualizaciones, se hace desde la vista de configuraciones de las técnicas y desde el diálogo de parámetros generales. También se puede manipular los datos interactuando directamente sobre algunas de las técnicas, haciendo uso de la vista de visualización.

3.4. Manejo masivo de información

Otro de los retos que se abordan en esta tesis, es el manejo masivo de información. En el epígrafe 2.7 se abordaron los principales formatos de datos científicos disponibles para el almacenamiento y visualización de grandes volúmenes de datos espacio-temporales.

En esta propuesta conceptual se recomienda la utilización de los formatos de datos HDF y NetCDF. Las principales razones para la selección de estos formatos de datos están dadas por la adecuación para la manipulación de datos espacio-temporales, la facilidad para integrarlos con sistemas de información geográfica y la disponibilidad de bibliotecas y funciones para su utilización. En particular la propuesta de utilización de estos formatos de datos científicos se puede analizar desde dos puntos de vistas: uno, el almacenamiento y gestión de los datos que se utilizan en la visualización de grandes volúmenes de datos espacio-temporales, que son tratados

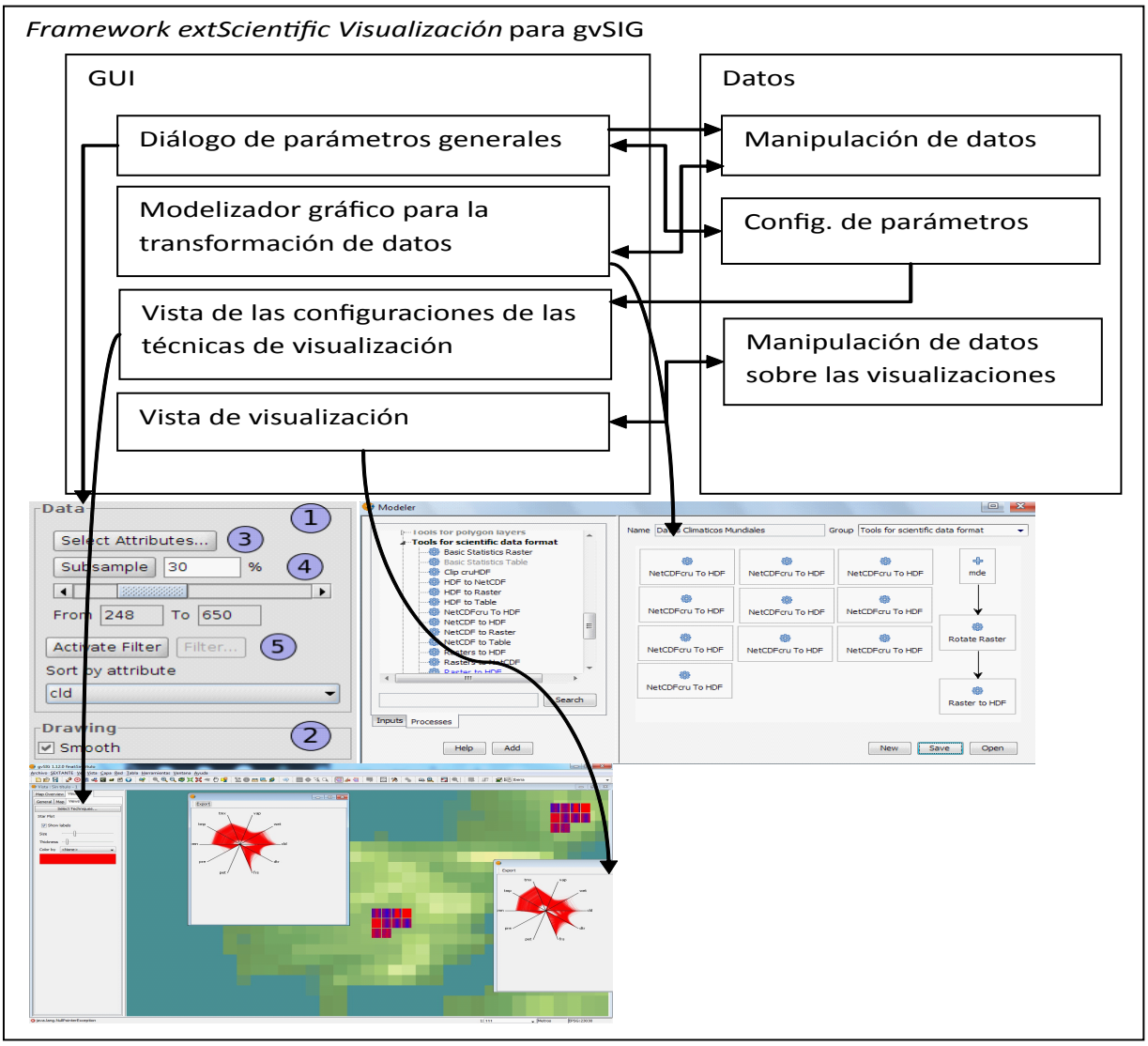

Figura 3.9 *Framework* extScientificVisualization para gvSIG.

en el capítulo 5, y el otro punto de vista tiene que ver con el desarrollo de herramientas para el soporte de estos formatos; temática abordada en el capítulo 6, donde se describen un conjunto de herramientas que permite la integración de estos formatos de datos con sistemas de información geográfica.

La figura 3.10 representa la interacción que existe entre los formatos de datos científicos y los diferentes módulos que se proponen para el análisis y manipulación de grandes volúmenes de datos. En este caso el módulo de visualización científica manipulará el formato HDF para lectura. Sin embargo, el módulo de manipulación de formatos de datos científicos utilizará el formato HDF para lectura y escritura. Este módulo podrá disponer del formato NetCDF para lectura y escritura. No se contempla la posibilidad de que se pueda escribir para un formato científico en el módulo de visualización científica.

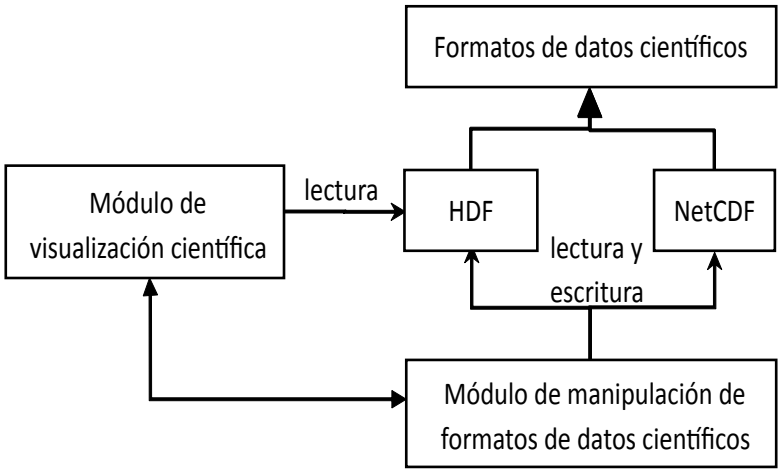

Figura 3.10 Propuesta de utilización de los formatos de datos científicos.

3.5. Tratamiento de datos y análisis multivariado

Con el análisis exploratorio de datos, se persigue integrar a los humanos en el proceso de exploración de datos, mediante la aplicación de las habilidades perceptuales de los humanos para analizar los grandes conjuntos de datos disponibles en los sistemas actuales. La idea básica de la exploración visual de datos, es presentar los datos de algunas formas visuales que permitan al usuario ganar conocimiento sobre la estructura de los datos, llegar a conclusiones e interactuar directamente con los datos (Keim *et al.*, 2005). Las técnicas de minería visual de datos han probado ser de gran utilidad para el análisis exploratorio de datos, y tienen un gran potencial para explorar grandes bases de datos. La exploración visual de datos es especialmente útil cuando se conoce poco acerca de los datos y el objetivo de la exploración es ambiguo. Dado que el usuario está involucrado directamente en el proceso de exploración, este puede ajustar y modificar el objetivo de la exploración.

La exploración visual de datos puede ser vista como un proceso de generación de hipótesis (Gahegan, 2005). Las representaciones visuales de los datos permiten al usuario conocer un poco más su estructura y descubrir nuevas hipótesis. La verificación de las hipótesis se puede confirmar a través de la visualización de datos, pero puede estar acompañada por el uso de técnicas automáticas de la estadística, reconocimiento de patrones y el aprendizaje automatizado. Además de la involucración directa del usuario en el proceso de análisis, el análisis exploratorio de datos tiene por encima de otros métodos automáticos de minería de datos, las ventajas siguientes:

- El análisis exploratorio de datos permite trabajar fácilmente con datos altamente densos y ruidosos.
- El análisis exploratorio de datos es intuitivo y requiere menos comprensión de complejos algoritmos estadísticos o matemáticos.
- La visualización puede suministrar una visión general cualitativa de los datos, permitiendo

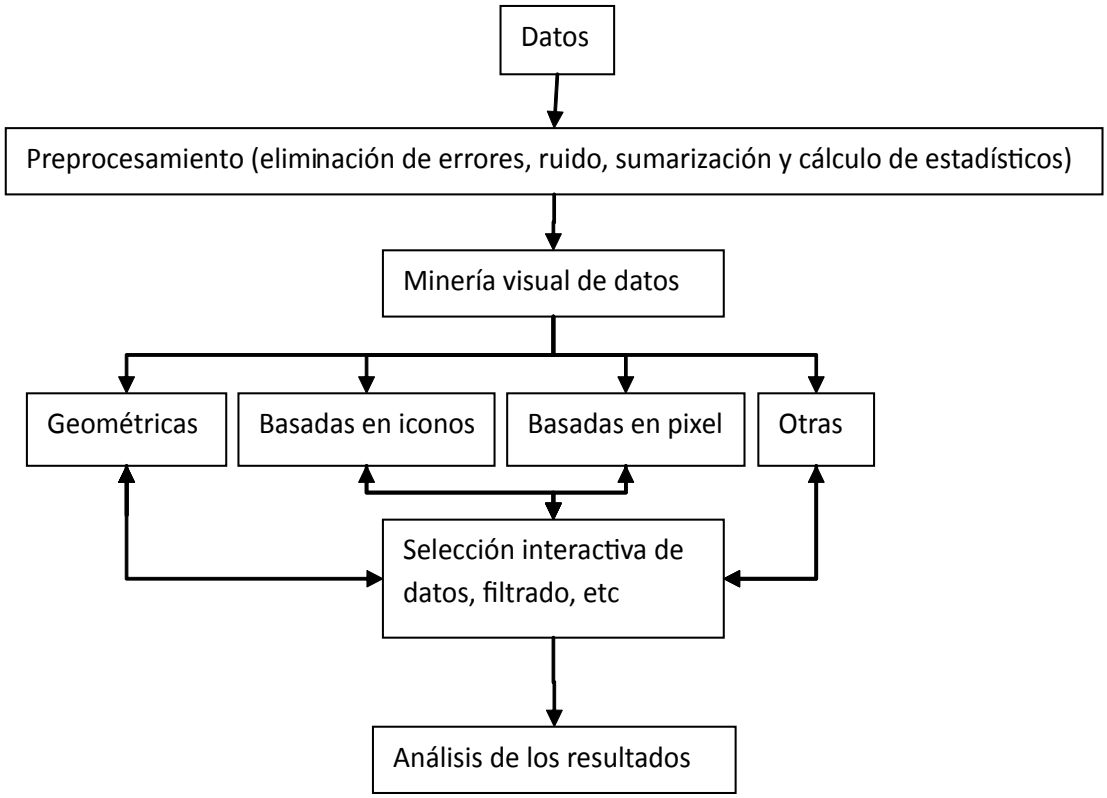

Figura 3.11 Esquema de procesamiento de los datos.

que algunos fenómenos sobre los datos sean aislados para realizar análisis cualitativos posteriores.

Sin embargo, el análisis exploratorio de datos incluye algunos procesamientos de datos como la reducción de dimensiones, la transformación de un espacio en otro, que normalmente requieren la comprensión de algunos conceptos básicos para entender mejor los algoritmos.

Como consecuencia, el análisis visual de datos puede servir para llegar a resultados superiores en menos tiempo, especialmente en casos en que los algoritmos automáticos fallan. Además, las técnicas de análisis exploratorio de datos suministran un alto grado de confianza en el descubrimiento. Este hecho ha provocado una alta demanda de técnicas de análisis exploratorio de datos, que son indispensables para su vinculación con técnicas automáticas de exploración.

La figura 3.11 muestra el esquema de procesamiento de los datos, que se propone con la minería visual de datos. Se parte de los datos crudos, que son preprocesados, en este paso se le eliminan errores y ruido, además es donde se realizan algunas sumarizaciones y cálculo de estadísticos básicos que luego serán utilizados por los métodos de visualización. Las visualizaciones pueden ser mediante técnicas geométricas, basadas en iconos, basadas en píxeles y otras de las que se analizaron en el epígrafe 2.3. El siguiente paso es la selección interactiva de datos y filtrado hasta llegar al análisis de los resultados.

3.6. Conclusiones parciales

En este capítulo se ha desarrollado un modelo que permite definir las pautas para integrar técnicas de visualización en sistemas de información geográfica que facilitan la manipulación e integración de datos geográficos y multiparamétricos. El modelo ha sido presentado mediante el uso de diagramas que abarcan la solución a los problemas planteados desde varias perspectivas. Se parte del esquema conceptual donde se muestran elementos generales para el diseño e implementación de herramientas que permiten el análisis exploratorio de grandes volúmenes de datos espacio-temporales. Se tuvieron en cuenta la visualización dinámica de este tipo de datos, el manejo masivo de información y el tratamiento y análisis multivariado.

Este modelo se ha llevado a cabo mediante el desarrollo de herramientas que se presentan en los capítulos 4, 5 y 6. El modelo puede ser generalizado y aplicado con herramientas de visualización, sistemas de información geográfica y formatos de datos científicos similares o diferentes de los seleccionados en esta tesis.

4 Análisis exploratorio de datos con baja densidad espacial

En el epígrafe 1.1 de esta tesis se planteó el problema general para el análisis exploratorio de datos con baja densidad espacial. En este capítulo, se retoman estos problemas y mediante la utilización del modelo propuesto se obtiene una herramienta que brinda una solución para el problema planteado. En el siguiente epígrafe se hace una introducción donde se plantea el contexto y las condicionantes para abordar este caso.

4.1. Introducción al análisis exploratorio de datos con baja densidad espacial

Gracias al desarrollo de las redes de comunicaciones, el *hardware*, la telefonía celular y el surgimiento de Internet se están generando a diario volúmenes de datos tan grandes y complejos, que no pueden ser analizados suficientemente en forma numérica. Se espera que para el 2020 existan más equipos que personas conectados a las redes, emitiendo y generando datos (Sundmaeker *et al.*, 2010). Los datos pueden estar relacionados con el estado de procesos, equipos o simplemente tomados por sensores o generados por simulaciones. El abaratamiento en el desarrollo de sensores, el *hardware* y los métodos de transmisión y comunicación de datos, está permitiendo que la captura y la generación de datos llegue a una gran cantidad de sectores como la industria, la sociedad, la ingeniería, el medio ambiente, etc, para resolver y analizar determinados problemas que existen actualmente.

Estos ejemplos que se han mencionado pueden generar datos multiparamétricos. Es también de esperar que la componente espacial asociada a los lugares donde fueron generados los datos, juegue un importante papel en el análisis espacial que se pueda realizar sobre los datos.

Los datos espacio-temporales con baja densidad espacial y amplios en el tiempo pueden ser encontrados en varias ramas de la ciencia, por ejemplo, datos históricos sobre estaciones meteorológicas donde se miden múltiples variables; otros ejemplos lo constituyen los censos tomados en regiones delimitadas por la distribución político-administrativa (municipios,

Tabla 4.1 Representación de los datos multiparamétricos

Atrib.1	**Atrib.2**	**...**	**Atrib.n**
$x_{1,1}$	$x_{1,2}$	...	$x_{1,n}$
$x_{2,1}$	$x_{2,2}$	...	$x_{2,n}$
...	...	...	...
$x_{m,1}$	$x_{m,2}$	...	$x_{m,n}$

provincias, países, etc.) que generalmente son pocas.

En este capítulo se trata el caso de manipular pocos conjuntos de datos multiparamétricos que pueden ser amplios en el tiempo. Es decir, desde el punto de vista espacial se dispone de un número finito y pequeño de conjuntos de datos multiparamétricos. Estos se desean analizar de manera independiente, sin tener en cuenta la ubicación, o de manera coordinada, manteniendo siempre una referencia a la ubicación geográfica sobre una capa vectorial. Se muestra un caso de estudio con datos meteorológicos de un reducido número de estaciones de medición. Los resultados presentados en este capítulo fueron publicados en la revista brasileña *Anuario do Instituto de Geociências* (Vázquez-Rodríguez *et al.*, 2013b), y presentados en la Conferencia Internacional de Descubrimiento de Conocimiento y Recuperación de Información, en Valencia, España (Vázquez-Rodríguez *et al.*, 2010d); ambas publicaciones indexadas por prestigiosas bases de datos bibliográficas.

4.2. Representación y estructura de los datos

La fuente de datos multiparamétricos consiste en un conjunto C de cardinalidad m en que cada elemento $O_i =< V_1..., V_n >, i = 1...m$ es una n-upla. A cada elemento de C se le conoce como instancia u observación, y cada componente de una observación es una dimensión o variable. En estos datos, cada variable puede tener un nivel de medición continuo o nominal.

Los conjuntos de datos multiparamétricos pueden tener una cantidad variable de dimensiones y observaciones. La tabla 4.1 muestra una representación de los datos multiparamétricos, algunas de las variables pueden ser consideradas como una clase.

Otros tipos de datos que se tienen en cuenta son los datos geográficos. Los datos geográficos para el caso tratado en este capítulo consisten en mapas, que se pueden utilizar como fondo para realizar un análisis teniendo en cuenta varias regiones a la vez. También pueden incluir las diferentes ubicaciones a las que pertenecen determinados conjuntos de datos multiparamétricos. Estos mapas de puntos se suministran en forma de mapa vectorial, como por ejemplo, en el formato *ESRI shape*.

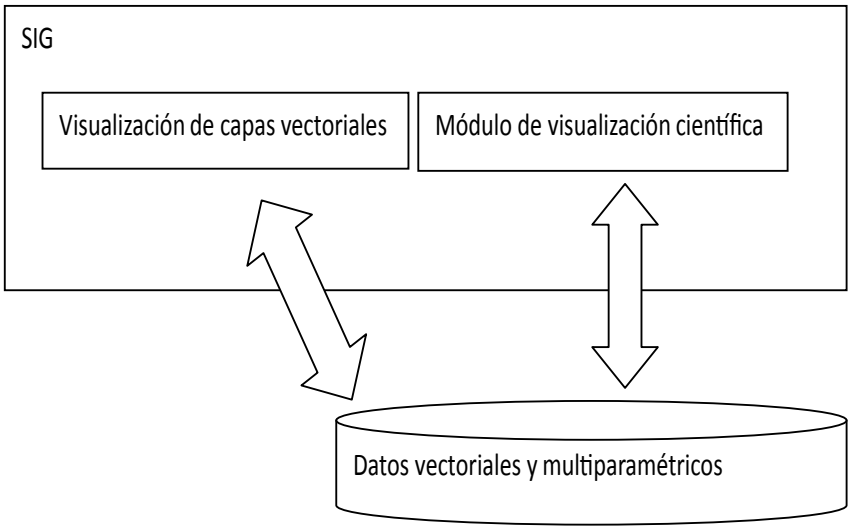

Figura 4.1 Arquitectura general para la solución propuesta para la baja densidad espacial.

4.3. Arquitectura general

La temática de la baja densidad espacial, desde el punto de vista de los formatos de datos de los sistemas de información geográfica, se puede abordar mejor desde la perspectiva de los datos vectoriales. La figura 4.1 muestra el esquema general que se propone para el análisis exploratorio de datos con baja densidad espacial.

Siguiendo la primera de las vías de la estrategia presentada en el modelo propuesto, la idea es tomar un núcleo de un sistema de información geográfica libre e incorporarle un módulo de visualización científica. Este módulo se encargará de manipular todas las técnicas de minería visual de datos. Se necesitará, además, modificar la funcionalidad de visualización de capas vectoriales del sistema de información geográfica para que soporte su interacción con las técnicas de visualización, que se puedan mostrar sobre los mapas. Los actores del modelo involucrados en esta integración son los desarrolladores de herramientas, quienes tendrán que tener en cuenta los principios del modelo para facilitar un conjunto de tareas a los actores especialistas, que utilizarán las herramientas.

El módulo de visualización científica basado en los principios, es el encargado de reunir las técnicas de visualización de datos multiparamétricos, y tiene que facilitar un conjunto de herramientas de *software* propias del análisis visual. Este módulo contiene dos paquetes: el primero tiene como objetivo la visualización de manera independiente de un único conjunto de datos multiparamétricos y el segundo la visualización de manera coordinada de varios conjuntos de datos multiparamétricos, teniendo una percepción espacial de la localización de cada conjunto de datos sobre un mapa. Estos constituyen los casos 1 y 2 presentados en la figura 3.8.

El objetivo de ambos paquetes es la visualización de datos multiparamétricos, por lo tanto las técnicas seleccionadas o implementadas pueden ser las mismas para cada caso. Entre las herramientas propias del análisis visual que se implementaron, en este caso se encuentran las

técnicas de visualización de datos multiparamétricos, la manipulación de vistas, las consultas y las herramientas de cálculo. Las técnicas que se decidió incluir en el módulo de visualización en este caso fueron:

- Coordenadas Paralelas.
- Gráfico de Andrews.
- Técnicas basadas en iconos:
  - Icono en forma de estrella.
  - Shapecoding.
  - Icono en forma de barras.
- Segmentos de Círculo.
- Patrones Recursivos.

La manipulación de vistas y las consultas sobre los datos se implementaron mediante paneles generales que permiten coordinar los parámetros de la técnica que esté seleccionada en un momento determinado. En el caso de la visualización sobre mapas se contempló la posibilidad de visualizar siempre con una misma técnica conjuntos de datos multiparamétricos asociados a diferentes regiones.

En el caso del sistema de información geográfica de base seleccionado es importante verificar que posee las funcionalidades básicas para manipular formatos de datos vectoriales, pues para el caso de la baja densidad espacial, la configuración de los conjuntos de datos multiparamétricos con la referencia espacial se facilita con mapas vectoriales de puntos y de polígonos.

4.4. Selección del modelo de datos y descripción del método de visualización

La integración del módulo de visualización de datos multiparamétricos se llevó a cabo utilizando la visualización independiente y la visualización coordinada. Estos dos enfoques ofrecen diferentes ventajas y desafíos en cuanto al desarrollo.

La visualización independiente significa visualizar, mediante técnicas de visualización científica, un único conjunto de datos. Este conjunto de datos puede corresponder a una única localización o incluir variables de diferentes localizaciones. En este caso la visualización se realiza normalizando los valores de las variables con respecto a sus extremos en el conjunto de datos.

La herramienta obtenida permite a través del módulo de visualización de datos multiparamétricos obtener conocimiento de datos medidos en iguales intervalos de tiempo con una gran cantidad de registros, pero pobres espacialmente. La solución consistió en integrar algunas técnicas de visualización de datos multiparamétricos en un sistema de información geográfica,

de tal manera que los datos pudieran ser analizados teniendo una percepción geográfica de su origen. Se implementó otra herramienta que permite al usuario preparar proyectos para ser utilizados en una visualización coordinada sobre mapas vectoriales.

Estos proyectos pueden ser personalizados para ser visualizados en un mapa vectorial de puntos (archivo *shape* dado por el usuario) o sobre un mapa vectorial de polígonos (archivo *shape* que se utiliza como fondo de la visualización). En este último caso, los gráficos se muestran sobre el centroide de la geometría poligonal correspondiente. Las técnicas que no se pueden visualizar como un gráfico sobre el mapa, debido a restricciones de espacio, se visualizan en paneles independientes, manteniendo una percepción geográfica del origen de los datos.

El sistema de información geográfica de escritorio seleccionado para ser extendido con el módulo de visualización científica fue gvSIG, el cual es un sistema de código abierto de la Generalitat Valenciana. Se trata de un sistema de información geográfica basado en Java que es muy fácil de extender y tiene muy buena documentación para desarrolladores. gvSIG tiene una jerarquía de clases bien estructurada y permite la lectura de varios formatos geográficos y no geográficos como tablas.

En la siguiente sección se explican las principales características del módulo que fue desarrollado e incorporado al sistema de información geográfica gvSIG, haciendo énfasis en los detalles correspondientes del modelo propuesto.

4.5. Caso de estudio: visualización de datos climáticos de la provincia de Villa Clara, Cuba

Para el análisis exploratorio de datos con baja densidad espacial se realizó un caso de estudio con datos meteorológicos. Los datos meteorológicos recopilados por el Instituto de Meteorología de la provincia de Villa Clara en Cuba representan una serie temporal de trece variables, medidas desde 1977 hasta la actualidad. Las variables meteorológicas recopiladas son: temperatura media, mínima y máxima promedio en la decena, humedad relativa media, mínima, y máxima promedio en la decena, déficit de saturación (promedio en grados en la decena), nubosidad (promedio en octavos en la decena), velocidad media del viento (promedio en la decena), lluvia total en la decena, insolación (promedio de horas luz en la decena), tensión de vapor de agua (promedio en la decena) y presión atmosférica (promedio en la decena).

Para cada una de las cuatro estaciones meteorológicas de la provincia, se cuenta con una serie temporal de las trece variables mencionadas. Las estaciones meteorológicas están localizadas en Yabú, Santo Domingo, Sagua la Grande y Caibarién.

Para el almacenamiento de los datos meteorológicos se han utilizado varios formatos como bases de datos, archivos en formato de texto (*arff*) y en formato de tablas (*dbf*). La tabla 4.2 muestra un fragmento de la forma en que se almacenan estos datos. Cada registro representa el promedio de los valores en diez días para cada una de las variables meteorológicas. Estos

Tabla 4.2 Estructura general de los datos meteorológicos

Año	Mes	Año - Decena	Temp	HumRel	...
1977	1	197701	20	73	...
1977	1	197702	18	77	...
...	...	...	...	...	...
1977	2	197705	23	83	...
...	...	...	...	...	...
1978	...	...	...	...	...
...	...	...	...	...	...

Tabla 4.3 Coordenadas geográficas de las estaciones meteorológicas

No.	Estación	Latitud	Longitud	Altura
1	Yabú 343	79.991	22.461	116.44
2	Sagua 338	80.092	22.806	12.06
3	Caibarién 348	79.471	22.497	46.27
4	Santo Domingo 326	80.226	22.586	45.35

registros se tienen para cada una de las estaciones. Los datos originales fueron procesados mediante métodos estadísticos y de limpieza de datos para corregirlos y eliminarles ruido.

Como se puede observar, existen varios rasgos relacionados con el tiempo (año, mes, decena y año-decena), por lo que se pueden lograr diferentes niveles de granularidad respecto al componente temporal. Esto sugiere la utilización de algunos de los principios del modelo como simplificar y resumir, dividir y agrupar.

Conjuntamente se posee la información cartográfica que incluye un mapa con la división político-administrativa en municipios de la provincia de Villa Clara, y un mapa de puntos con las coordenadas de las localizaciones de las estaciones meteorológicas. La tabla 4.3 muestra las coordenadas geográficas de las estaciones.

Mediante el asistente de configuración de proyectos de visualización coordinada, se configuró un proyecto en el cual se tomó como mapa base el mapa de la provincia Villa Clara y para las localizaciones se empleó un mapa de puntos con las localizaciones de las estaciones meteorológicas, haciendo corresponder los archivos de datos multiparamétricos a las localizaciones.

Otra vía para analizar los datos sería realizar la visualización de cada conjunto de datos por separado mediante la extensión para la visualización independiente. En ese caso el análisis sería localizado para cada estación meteorológica pues la visualización depende únicamente de los valores del conjunto de datos que utiliza.

Preparación de los datos para ser visualizados

Para visualizar con una visualización coordinada en gvSIG, son necesarios 2 tipos de datos: datos geográficos y datos multiparamétricos. Los datos geográficos pueden incluir un mapa de

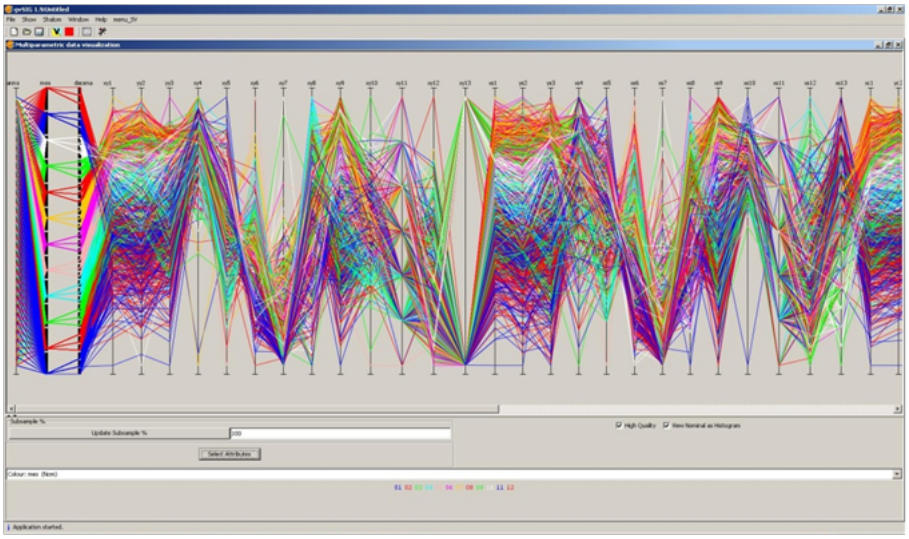

Figura 4.2 Visualización independiente de todas las variables mediante la técnica de coordenadas paralelas.

fondo y un mapa de puntos, o sólo un mapa de fondo; ambos como archivos de formatos de datos vectoriales. Cada punto de datos debe coincidir con un archivo de datos multiparamétricos.

Esta extensión permite leer los datos multiparamétricos en el formato de tabla o de archivos planos separados por coma, que especifican los metadatos como nombres y tipos de todas las variables de la tabla. Los datos de cada lugar puntual tienen que coincidir con el número de variables y registros. Se admiten valores ausentes. En este aspecto tienen un papel importante las tareas de manipulación de datos y de cálculo, pues en ocasiones se necesita preparar los datos para hacer sumarizaciones y simplificaciones con vistas a realizar posteriormente las visualizaciones.

Visualización

Existen varias formas de visualizar datos con técnicas de visualización de datos multiparamétricos. Algunas de ellas permiten visualizaciones en mapas, donde se percibe a la asociación de los datos con la zona geográfica. En particular, las técnicas de visualización de datos multiparamétricos no necesariamente tienen que estar asociadas a un mapa, como se muestra en la figura 4.2. Sin embargo, estas técnicas se pueden aplicar a todas las variables en cada punto de datos para obtener correlaciones entre ciertas variables de varios puntos. Es posible utilizar, por ejemplo, coordenadas paralelas para mostrar las $13 * 4 = 52$ variables correspondientes a las 4 estaciones meteorológicas. Aquí se tendrían en cuenta tareas tanto elementales como sinópticas para analizar todas las variables a la vez y poder descubrir patrones y relaciones entre varias de ellas. Este pudiera ser un primer paso para aplicar el principio de buscar lo reconocible para luego enfocarse en otros principios, haciendo uso de otros tipos de tareas. En este punto se puede interactuar con varias tareas del tipo “muestra”.

Otra variante de aplicación es mostrar los datos de cada punto separadamente, y mostrar un mapa donde sea evidente a qué región corresponde cada gráfica. Esto se puede lograr con

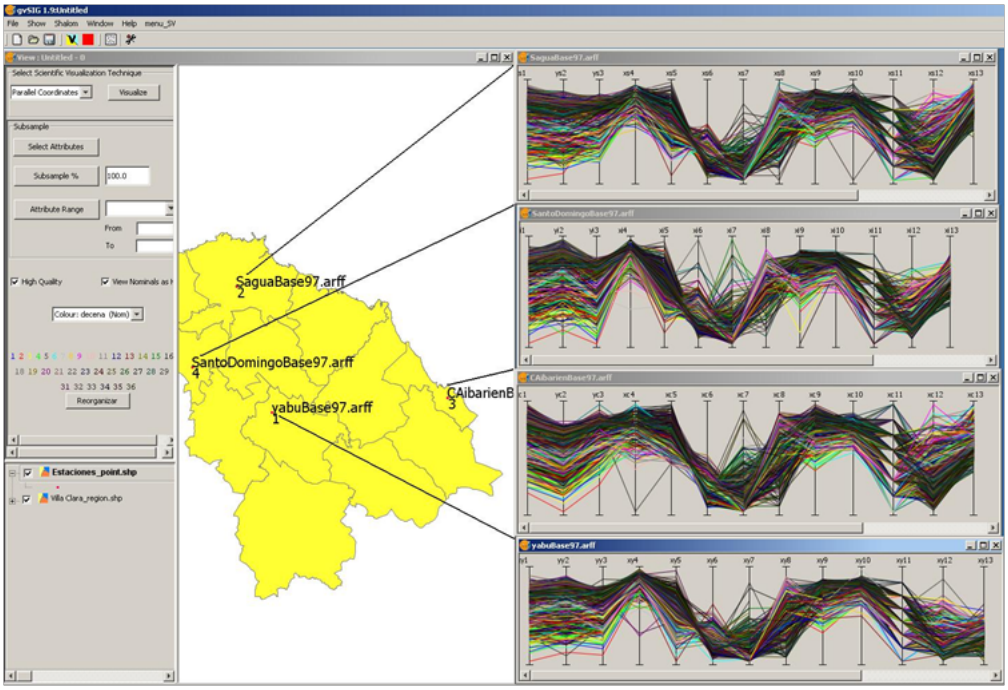

Figura 4.3 Visualización coordinada mediante coordenadas paralelas.

una flecha que conecta la región en el mapa con el panel de visualización o nombrando cada panel con el mismo nombre del archivo de punto de datos visualizado en el mapa (obsérvese la figura 4.3). Aquí se evidencia el enlace de múltiples vistas coordinadas por herramientas de configuración de parámetros.

Algunas técnicas de visualización de datos multiparamétricos, se pueden presentar directamente sobre el mapa. Algunos ejemplos de esas técnicas son segmentos de círculos, patrones recursivos y técnicas basadas en iconos (obsérvense las figuras 4.4 y 4.5). Con las técnicas basadas en iconos es posible mostrar un icono para cada fuente de datos (una estación meteorológica en nuestro caso), que represente el conjunto de variables para una observación en un momento dado. Se pueden utilizar algunos mecanismos para desplazarse en el tiempo (controles deslizantes, barras de desplazamiento), lo que permite cambiar el icono sobre el mapa en consecuencia con las acciones realizadas (obsérvese la figura 4.5). Esto permite al usuario estudiar la evolución de los datos en el tiempo. Este mismo mecanismo se puede utilizar para visualizar los datos en relación con otras variables.

En los paneles de la izquierda se pueden observar controles deslizantes que permiten modificar algunos parámetros de las imágenes. Entre otros, se puede modificar el tamaño de los gráficos, el tamaño del píxel, las etiquetas de las variables. De esta forma se cumplen algunos principios como ampliar, enfocar y atender a particulares para ver mejor las relaciones. Existe la posibilidad de mostrar la leyenda de una visualización determinada para mantener una correspondencia entre los colores y los elementos de datos que se están interpretando. Algunas técnicas como las basadas en iconos, cuando se muestran sobre el mapa, se pueden interpretar mejor mediante animaciones, y, en consecuencia, esta es una opción que se brinda en este módulo.

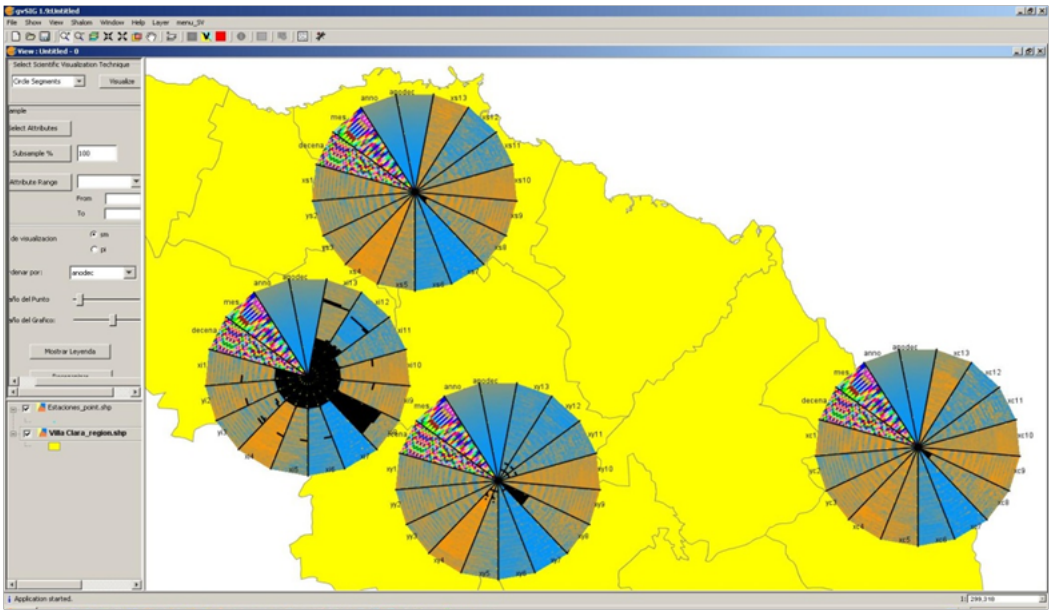

Figura 4.4 Visualización coordinada con segmentos de círculo sobre el mapa.

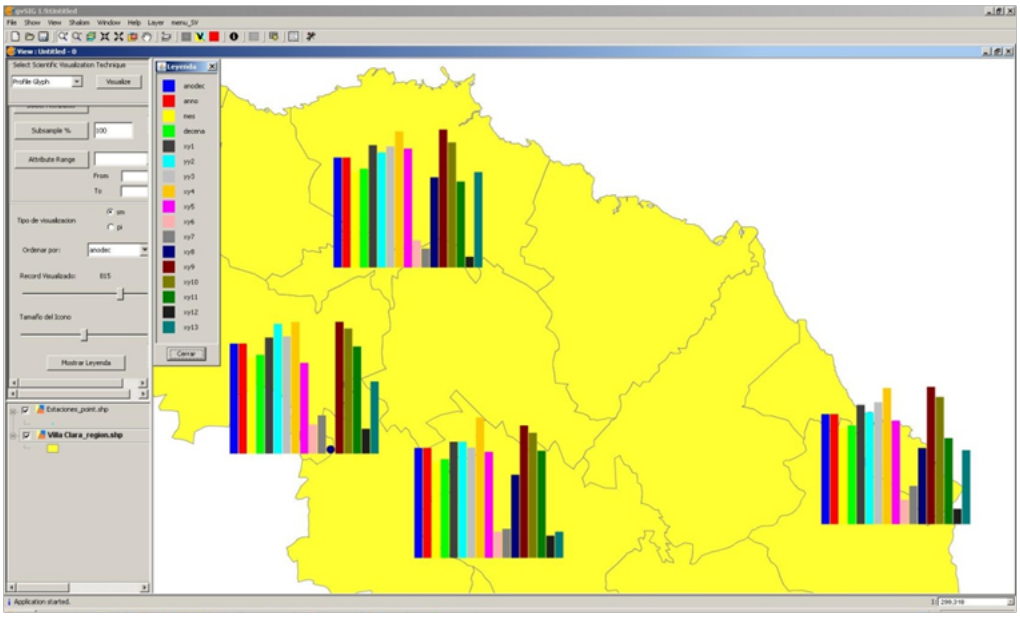

Figura 4.5 Visualización coordinada con *Profile glyphs* sobre el mapa.

Todos estos métodos se aplicaron en el módulo desarrollado para gvSIG. Por ejemplo, uno de los métodos desarrollados es el de visualizaciones no coordinadas, el cual es muy útil cuando los datos no están asociados con un mapa. En estos casos el sistema permite al usuario cargar un archivo de datos para ser analizado usando alguna de las técnicas siguientes:

- Coordenadas paralelas.
- Gráficos de Andrews.
- Campo de estrellas.
- Iconos en forma de barras.
- Shapecoding.
- Segmentos de círculo.
- Patrones recursivos.

Para la visualización coordinada sobre el mapa se incluyen las técnicas siguientes:

- Coordenadas paralelas (en paneles independientes).
- Gráficos de Andrews (en paneles independientes).
- Campo de estrellas (un registro a la vez).
- Shapecoding (un registro a la vez).
- Iconos en forma de barras (un registro a la vez).
- Segmentos de círculo (todos los registros seleccionados).
- Patrones recursivos (todos los registros seleccionados).

En nuestro caso de estudio, se obtuvieron los mejores resultados con la técnica de patrones recursivos. Esta es una excelente técnica para llevar a cabo un análisis espacio-temporal en sistemas de información geográfica. El módulo desarrollado permite algunas funcionalidades que están disponibles para todas las técnicas:

- Selección de atributos (sólo los atributos seleccionados se muestran en el gráfico). Aquí se involucran tareas de manipulación de datos y herramientas de consulta.
- Selección de un porcentaje de los registros. Esta funcionalidad permite establecer estructura, pues selecciona un subconjunto de referencias al igual que la siguiente funcionalidad.
- Selección de acuerdo con un rango de valores de un atributo (hace un sub-ejemplo de los datos en un rango de valores seleccionados para un atributo dado).
- Se muestra la leyenda (muestra la gama de colores global para cada variable, que toma los valores mínimo y máximo de todos los conjuntos de datos, a los valores nominales se les asigna un color diferente para cada valor, algunas técnicas como iconos en forma de barras muestran en la leyenda un color diferente para cada atributo). Este aspecto ayuda a establecer referencias para involucrar conocimiento sobre el dominio.

- La reorganización de atributos (se establece un nuevo orden de los atributos). Esta funcionalidad contribuye al principio que posibilita ver relaciones y permite establecer estructura.

La opción de reorganización de los atributos ordena todos los conjuntos de datos de acuerdo con un atributo dado. Las técnicas basadas en píxeles y técnicas basadas en iconos utilizan esta funcionalidad para ordenar todos los valores según este atributo. Al utilizar el tiempo como atributo, el usuario puede analizar los datos en el tiempo.

Cada técnica tiene su panel de configuración y las características propias que se pueden configurar por separado. Además existe un grupo de aspectos generales relacionados con los tipos de técnica. Por ejemplo, para todas las técnicas basadas en iconos visualizados sobre el mapa, el usuario puede cambiar el tamaño de los iconos para obtener una mejor visualización, y se puede utilizar una barra para desplazarse por los registros usando un atributo seleccionado, es decir, un atributo de tiempo. Mediante esta opción de desplazamiento se puede hacer uso del principio de atender a particulares.

Todas las técnicas desarrolladas basadas en píxeles también permiten modificar el tamaño de los gráficos. Estas técnicas se pueden visualizar tanto en paneles independientes, como sobre el mapa. Muchas de ellas permiten hacer uso de los principios ver el todo, ampliar y enfocar y establecer estructura.

Las técnicas geométricas desarrolladas fueron diseñadas para ser visualizadas en paneles independientes, uno para cada punto de datos. También son coordinadas por el panel de configuración principal. Aquí se evidencian las herramientas de manipulación de vistas.

La técnica basada en píxeles patrón recursivo, tiene una manera especial de interactuar con datos, que permite definir diferentes niveles recursivos. Esto brinda la posibilidad de organizar los datos en diferentes formas, sobre todo cuando tienen un orden natural de acuerdo con una dimensión (por ejemplo, los datos de series de tiempo).

Una posibilidad simple es la de organizar los elementos de datos de izquierda a derecha línea por línea. Otra posibilidad es la de organizar los píxeles columna por columna. Una manera de mejorar la visualización es la organización de los píxeles en pequeños grupos y organizar los grupos para formar un patrón global. Esta estrategia corresponde a un enfoque en dos fases con un patrón de primer orden formado por la agrupación de los píxeles y con un modelo de segundo orden formando la organización global. Al tomar los resultados de la estructura de segundo orden como el elemento básico de construcción de una estructura de tercer nivel, se puede introducir un patrón de tercer orden. Este proceso puede ser repetido hasta un nivel arbitrario, formando un esquema general recursivo. Esta técnica demuestra un ejemplo clásico de utilización del principio establecer estructura.

La posibilidad de ver el todo, en este caso permite adquirir una idea general del comportamiento global de los datos. Este principio puede ser utilizado para detectar donde hay valores ausentes y comparar múltiples variables en los diferentes momentos de tiempo.

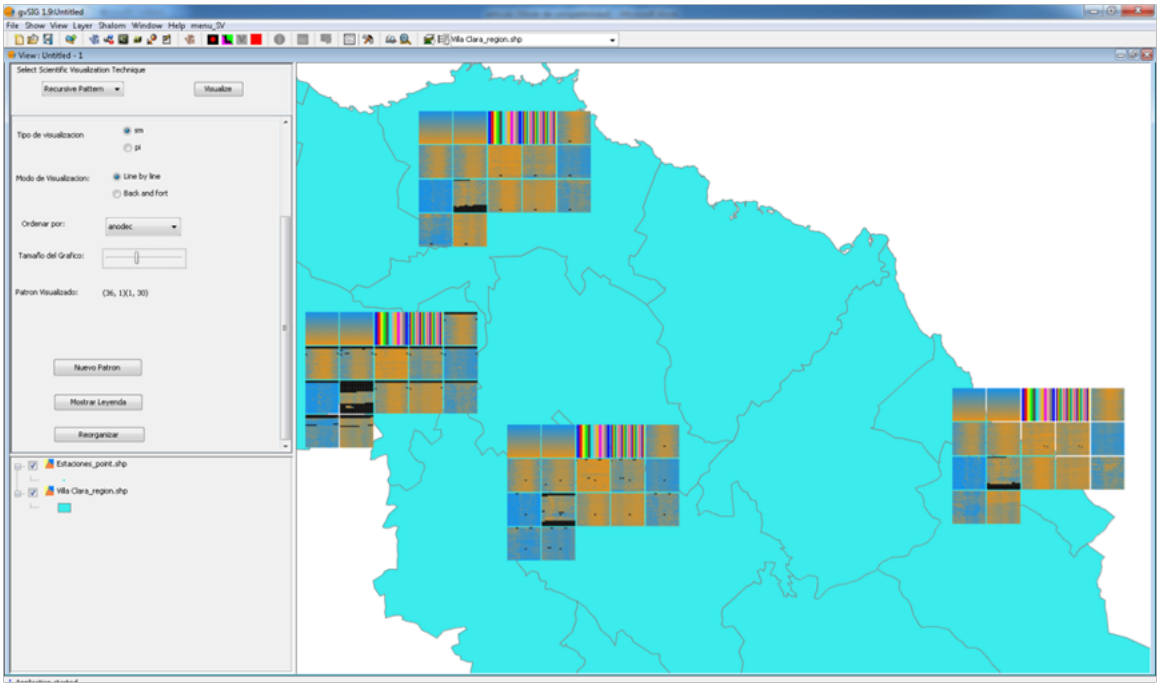

Figura 4.6 Se muestra la visualización coordinada con patrones recursivos. Cada punto de datos muestra 17 atributos, es decir, los atributos de tiempo como el año, mes, decena, año decena, y las 13 variables de información. El patrón recursivo utilizado en la imagen es (36,1) (1,30), que muestra un año por cada fila de izquierda a derecha (las 36 decenas de un año), y 30 años de arriba a abajo, se muestran 1080 valores para cada atributo (cuadrados pequeños).

La integración de las técnicas de visualización de datos multiparamétricos en gvSIG, se llevó a cabo mediante la incorporación de 2 extensiones a gvSIG, una para la visualización independiente y otra para la visualización coordinada; esto se realizó creando clases que heredan de la clase Extension de gvSIG y siguiendo la metodología documentada. A través de ese enlace el módulo de visualización, con toda su jerarquía de clases, queda incorporado a gvSIG y listo para utilizarse. Es importante destacar que la visualización coordinada es la más adecuada para realizar los análisis espacio-temporales más complejos. Una explicación más detallada del uso de este módulo se puede encontrar en los anexos 1 y 2.

4.6. Conclusiones parciales

El análisis exploratorio de datos con baja densidad espacial fue tratado en este capítulo como un caso especial que permite analizar visualmente datos multiparamétricos de manera independiente o coordinada asociados a pocos lugares del espacio. Haciendo uso del modelo propuesto en el capítulo 3, se obtuvo una herramienta con múltiples técnicas de visualización de datos multiparamétricos que fueron integradas como una extensión de la versión 1.9 de gvSIG.

Se presentó un caso de estudio con datos climáticos de la provincia de Villa Clara, Cuba

y los principales resultados fueron publicados en la revista brasileña *Anuario do Instituto de Geociências* (Vázquez-Rodríguez *et al.*, 2013b) y presentados en un evento internacional de prestigio (Vázquez-Rodríguez *et al.*, 2010a). Estas herramientas pueden ser generalizadas y utilizadas en múltiples áreas de aplicación debido a que son generales y pueden ser configuradas para múltiples propósitos dentro del análisis exploratorio de datos.

5 Análisis exploratorio de datos con alta densidad espacial

El problema general para el análisis exploratorio de datos con alta densidad espacial fue presentado en el epígrafe 1.1. En el siguiente epígrafe de este capítulo se retoman estos problemas, se introducen las condicionantes y se establece el contexto en que se enmarca esta investigación. Se ha utilizado el modelo propuesto para obtener una herramienta para solucionar los problemas del análisis exploratorio de datos con alta densidad espacial. La vía de la estrategia seguida en este caso es la de partir de un núcleo de un sistema de información geográfica y añadirle las funcionalidades necesarias para solucionar el problema en cuestión. En el epígrafe 5.2 se muestra la estructura y representación de los datos para el caso tratado en este capítulo. La arquitectura general propuesta y la solución se plantean en los epígrafes 5.3, 5.4 y 5.5.

5.1. Introducción al análisis exploratorio de datos con alta densidad espacial

A menudo la información de que disponemos no se encuentra presente en todo el espacio, sino en lugares puntuales, como los ejemplos mostrados en el capítulo anterior. Esa información puntual puede ser amplia en el tiempo y similar en estructura a la información disponible en varios lugares puntuales. En la actualidad existe la posibilidad de obtener información temporal de múltiples variables, que si bien se obtienn en lugares puntuales donde se encuentran las estaciones de medición, mediante métodos de interpolación espacial se puede generar información estimada de los lugares que no fueron muestreados, obteniéndose de esta forma valores suficientemente realistas de los lugares no muestreados.

Gracias al desarrollo de la aeronáutica se ha logrado desplegar una gran cantidad de sensores remotos que recogen información de toda la superficie del planeta. De esta manera se crean secuencias temporales de múltiples variables. La climatología es una de las áreas más avanzadas en la captura y generación de este tipo de información densa desde el punto de vista espacial y temporal.

Los datos de series temporales a menudo se generan por la monitorización de procesos. El análisis de series temporales se lleva a cabo cuando los puntos de datos obtenidos en el tiempo pueden tener una estructura interna, como una autocorrelación o variación estacional.

$$X = \{X_1, X_2, \ldots\} = \{X_t : t\ pertenece\ a\ T\}$$

En las ciencias geoespaciales pueden encontrarse comúnmente las series temporales, donde los datos pueden ser representados utilizando mapas (Yuan, 1996). De esta manera los sistemas de información geográfica han ido incorporando operaciones para manipular series temporales, como por ejemplo la nueva versión de GRASS GIS (Neteler *et al.*, 2012). En este caso los X_i de la ecuación anterior son mapas.

Cuando se analizan grandes volúmenes de datos en series temporales, es común encontrarse con varios problemas: las secuencias no pueden ser analizadas con simples imágenes de una variable, debido a que se requiere tener en cuenta el tiempo, y el solapamiento de todos los mapas provoca que se tengan que analizar en secuencias de imágenes o animaciones. Los métodos y herramientas actuales no permiten el análisis visual espacio-temporal de múltiples variables a la vez.

El objetivo de este capítulo no es manipular series de mapas, sino suministrar herramientas que permitan interactivamente explorar toda esta información y descubrir aspectos relacionados con su estructura interna y su relación con otras variables. Esto es especialmente útil cuando no se tiene conocimiento sobre estos aspectos en los datos.

En este capítulo se trata el caso de manipular una gran cantidad de conjuntos de datos multiparamétricos que pueden ser también amplios en el tiempo. Es decir, desde el punto de vista espacial se dispone de un número alto de conjuntos de datos multiparamétricos que están distribuidos en forma de mallas regulares. Estas mallas regulares se utilizan para simular fenómenos continuos espacialmente, es por eso que de los formatos de datos de los sistemas de información geográfica, el formato *raster* es el que más se ajusta para manipular la información que se describe en este capítulo. Aunque se intenta simular fenómenos continuos espacialmente, en la práctica también se dispone de un número finito de conjuntos de datos multiparamétricos, pero mucho mayor que los tratados en el capítulo anterior.

Este gran volumen de información trae consigo nuevos retos computacionales para gestionar, manipular y visualizar múltiples variables espacio-temporales.

Los resultados presentados en este capítulo fueron publicados en la *Revista Técnica de la Facultad de Ingeniería de la Universidad del Zulia* (Vázquez-Rodríguez *et al.*, 2015). Revista que figura en el listado de revistas del *Journal Citation Report*.

En esta investigación se realiza la visualización de series temporales de múltiples variables, mediante la utilización de técnicas de visualización de datos multiparamétricos integradas dentro de un sistema de información geográfica. La estructura de los datos requerida para esta investigación se describe en el siguiente epígrafe. El uso de las herramientas desarrolladas en

Tabla 5.1 Características del *dataset* con los datos climáticos mundiales

Dataset	Espacio	Tiempo	Variedad	Variables
CRU TS3.21	0.5° global	1901-2012	Series temporales de observación	pre, tmp, tmx, tmn, dtr, vap, cld, wet, frs, pet

esta investigación posibilitó presentar un caso de estudio con estos datos donde se demuestra la efectividad de las técnicas para extraer tendencias, correlaciones y patrones.

5.2. Representación y estructura de los datos

Los datos utilizados para este caso han sido tomados por el instituto *Climatic Research Unit* (CRU) de la universidad de East Anglia en Reino Unido. Este instituto es reconocido como uno de los líderes mundiales en el estudio concerniente al cambio climático natural y antropogénico. La unidad ha desarrollado un número de *datasets* que son ampliamente usados en investigaciones climáticas, incluyendo los registros de temperaturas globales que son usados para monitorear el estado del clima.

El objetivo fundamental de la unidad de investigaciones climáticas es mejorar el conocimiento científico en tres áreas:

- Historia del clima pasado y su impacto sobre la humanidad.
- El curso y las causas del cambio climático durante el siglo actual.
- Perspectivas para el futuro.

Los *datasets* son manipulados por una variedad de personas y proyectos dentro del CRU. Algunos están disponibles *on-line*, otros pueden ser obtenidos a través de las personas responsables de estos. Los archivos terminados en ".gz" pueden ser descomprimidos usando gzip en muchas plataformas o jzip en Windows.

Características de los datos de origen

Como se mencionó anteriormente, el CRU dispone de varios *datasets* con diferentes características. Para ejemplificar la solución propuesta para este caso, se ha montado un caso de estudio con el *dataset* CRU TS 3.21 que almacena los datos en forma de cuadrícula de alta resolución. En la tabla 5.1 se muestran algunas de sus propiedades.

En este tipo de *dataset* los datos climáticos son tomados durante meses sobre el último siglo, y su principal propósito es permitir que las variaciones en el clima se puedan comparar con las variaciones en otros fenómenos.

Tabla 5.2 Variables climatológicas contenidas en el *dataset*

Etiqueta	Variable	Unidad de medida
cld	Cobertura nubosa	Porcentaje (%)
dtr	Rango de temperatura diurna	Grados Celsius
frs	Frecuencia de días con escarcha	Días
pet	Potencial de evapotranspiración	Milímetros (mm)
pre	Precipitación	Milímetros (mm)
tmp	Temperatura media diaria	Grados Celsius
tmn	Promedio mensual de la temperatura mínima diaria	Grados Celsius
tmx	Promedio mensual de la temperatura máxima diaria	Grados Celsius
vap	Presión de vapor	Hectopascal (hPa)
wet	Frecuencia de días húmedos	Días

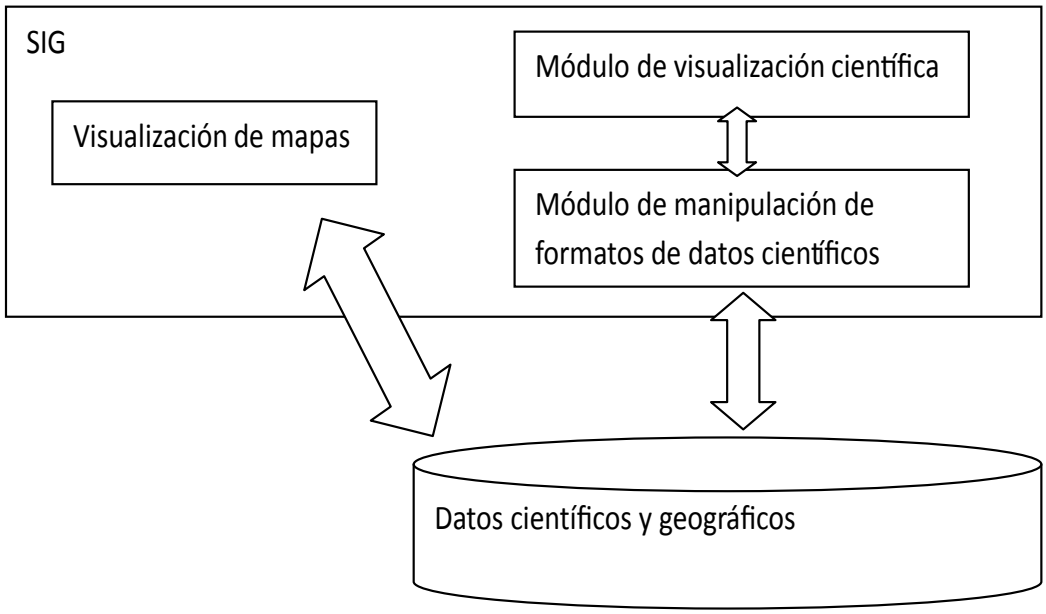

Figura 5.1 Arquitectura general.

En la tabla 5.2 se observan algunas abreviaciones usadas para referirse a las variables climatológicas.

5.3. Arquitectura general e interacción del usuario con el *software*

Los sistemas de información geográfica no están diseñados para trabajar de forma óptima con grandes conjuntos de información, por lo tanto se quiere modificar la forma en que se cargan los datos para poder ser utilizados por la extensión para la visualización.

La figura 5.1, muestra la arquitectura propuesta, en la cual se implementaron dos nuevos

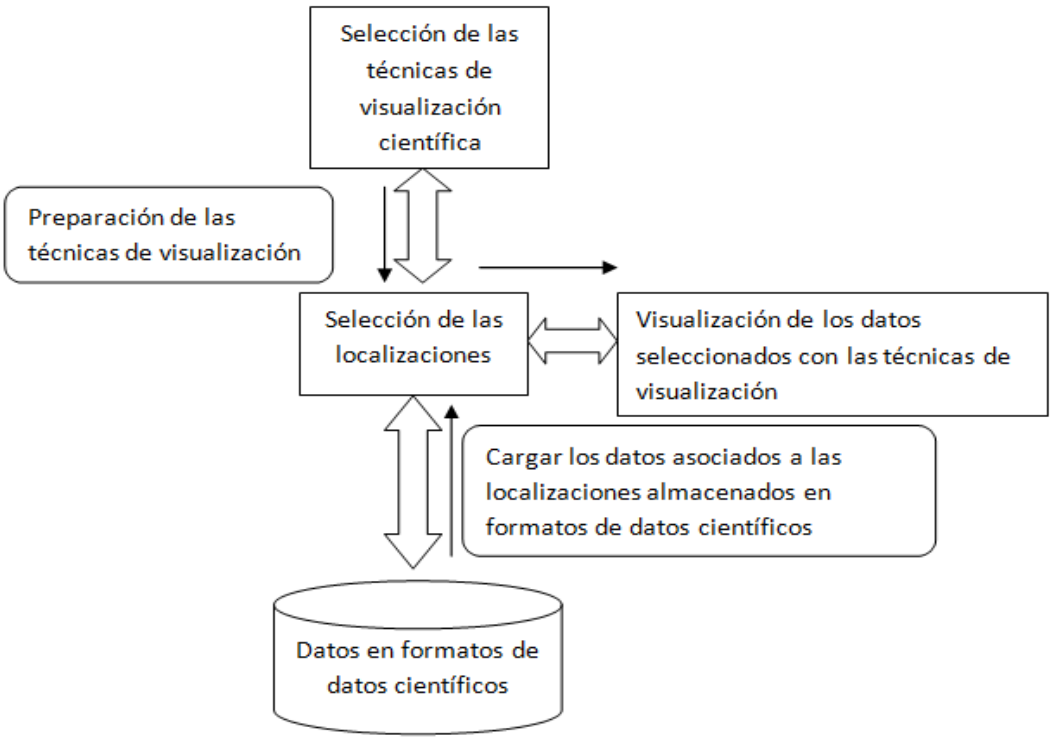

Figura 5.2 Interacción del usuario con el *software*.

módulos al sistema de información geográfica: el módulo de visualización científica con todas las posibilidades de mostrar diferentes técnicas e interactuar con ellas, y el módulo para la lectura de datos con formatos de datos científicos que permite manejar grandes volúmenes de datos, como por ejemplo HDF.

Un actor analista tiene la posibilidad de seleccionar las técnicas de visualización que desea en un análisis, para luego mostrar los datos asociados a las regiones del mapa que sean de su interés. Además permite escoger cuál de las técnicas seleccionadas será mostrada sobre el mapa de fondo y cuáles estarán en paneles independientes. Este es el estado de preparación de las técnicas de visualización. Una vez seleccionadas las localizaciones se procede a analizar los datos mediante las visualizaciones. Los datos son cargados directamente de los formatos de datos científicos. Esta arquitectura permite mostrar con las mismas técnicas de visualización, los datos asociados a distintas localidades del mapa y así poder analizar los resultados para extraer conclusiones o corroborar hipótesis. Para una mejor comprensión, en la figura 5.2 se muestra cómo es la interacción del usuario con el *software*.

5.4. Selección del modelo de datos y descripción del método de visualización

Después de analizar los modelos descritos en Andrienko y Andrienko (2006) y en el epígrafe 2.5.1, el modelo de datos seleccionado para este caso utiliza como referencia el espacio y el

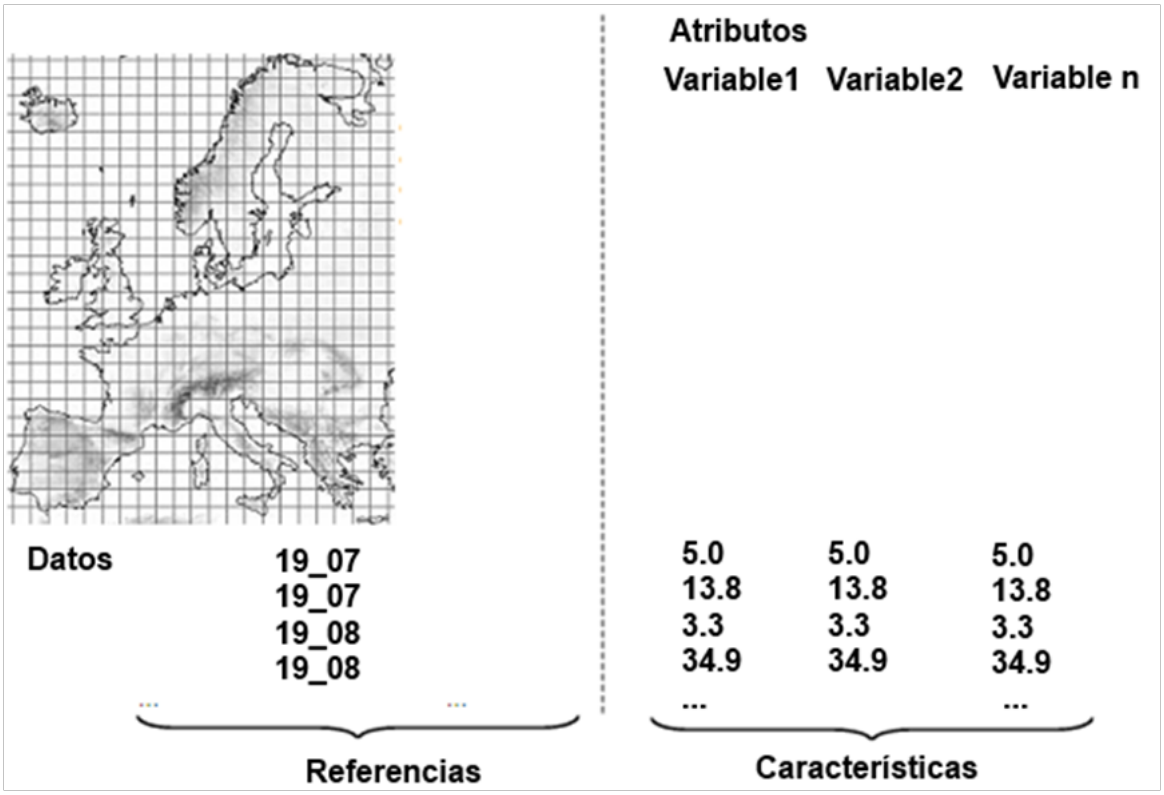

Figura 5.3 Modelo de datos seleccionado para el problema tratado en este capítulo. Modificado de (Andrienko y Andrienko, 2006).

tiempo, y el resto de las variables como las características, aunque el tiempo puede ser visto como otra variable de las características. La figura 5.3, muestra un ejemplo del modelo de datos seleccionado para este enfoque.

A continuación en la figura 5.4, se muestran los estados por los que pasa el sistema durante el proceso de visualización coordinada.

Inicialmente el sistema pasa al estado *cargando proyecto de visualización científica*, en el cual se carga el mapa de fondo y se abre el archivo HDF que contiene todos los datos que van a ser analizados. De este estado se pasa a *seleccionando técnicas de visualización científica*, momento en el cual hay que seleccionar las técnicas que se desea visualizar para un análisis determinado, especificando cuál de ellas se mostrará sobre el mapa y cuáles en paneles independientes.

Luego se pasa al estado *seleccionando localizaciones y mostrando representación visual de los datos*; durante este estado se pueden seleccionar los lugares del mapa de fondo de donde desea cargar los datos que estarán involucrados en el análisis. En cada selección se estarán leyendo los datos correspondientes desde el archivo del formato de dato científico, y se pasa al estado *cargar datos de archivo HDF*.

El proceso de lectura de los datos dada una petición en una celda de la malla es el siguiente: abrir el conjunto de datos que almacena los índices (Matriz de M por N) y cargarlo en memoria. Al acceder a esta matriz en (fila, columna), se devuelve el número de la celda si contiene

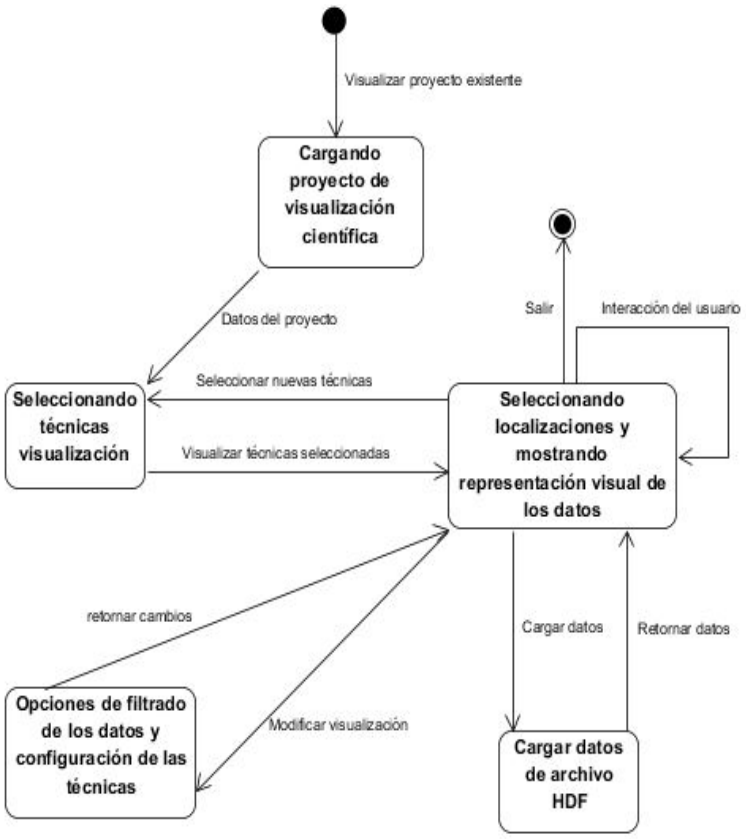

Figura 5.4 Diagrama de transición de estados para visualizar de forma coordinada.

información. Si no contiene información, el valor en el escaque dado es 0. Luego se abre cada uno de los conjuntos de datos que almacenan las variables climáticas y se lee sólo los Q valores de cada conjunto de datos, accediendo por las filas al número de celda devuelto restándole 1 y leyendo en cada conjunto de datos los Q valores de la fila correspondiente. De esta forma se seleccionan $V * Q$ valores. Además se lee el mínimo y máximo global de cada variable.

Luego se pasa nuevamente al estado *seleccionando localizaciones y mostrando representación visual de los datos*; en este paso, la información obtenida en el estado anterior, se le pasa a cada técnica para que se visualice sobre la celda correspondiente en el mapa de fondo o en paneles independientes, en dependencia de la configuración seleccionada.

Al interactuar con otras localizaciones, se repite el proceso para esa nueva celda. Se pueden modificar los parámetros de las visualizaciones, pasando al estado *opciones de filtrado de los fatos y configuración de las técnicas*. Durante este estado, se pueden seleccionar las variables que se quieren visualizar, el porcentaje de los datos que se quiere mostrar, seleccionar de acuerdo con un rango de valores de un atributo (hace un subejemplo de los datos en un rango de valores seleccionados para un atributo dado), entre otras opciones.

Además, cada técnica tiene un panel de configuración independiente, que permite personalizar las características específicas de cada una de ellas. Luego de realizar los cambios deseados, se retorna al estado *seleccionando localizaciones y mostrando representación visual de los datos* para mostrar los cambios realizados. Es aquí donde se interactúa con las técnicas.

Visualización

En esta sección se realiza una presentación de las principales funcionalidades de la herramienta obtenida para solucionar el problema en este caso. Se efectúa un análisis detallado de las opciones y modos de uso de algunas de las formas de visualización implementadas, que constituye una guía para explotar los beneficios de la extensión realizada.

Para utilizar la extensión desarrollada basta copiar el módulo compilado en el directorio de extensiones de gvSIG. Se recomienda utilizar la versión 1.12 de gvSIG. Esta fue la versión que se utilizó para todas las pruebas y análisis presentados en este capítulo.

La extensión realizada a gvSIG para este caso, permite la exploración y análisis visual de datos mediante las técnicas: coordenadas paralelas, gráfico de Andrews, segmentos de círculo, patrones recursivos, caras de Chernoff, figuras con palillos, matriz de diagramas de dispersión, *table lens*, rueda de tiempo, combo temporal, coordenadas de estrella, espiral de tiempo, río temático, vista circular y mapas auto-organizados. Permite visualizar coordinadamente varios conjuntos de datos, con diferentes técnicas a la vez. Los atributos y parámetros de las visualizaciones se pueden controlar de manera global o por tipos de técnicas.

El actor analista tiene la posibilidad de realizar el análisis visual de varios conjuntos de datos de forma coordinada, con una percepción del origen de los datos que están siendo visualizados referenciando las técnicas a un mapa. Para realizar esta operación es necesario tener un proyecto de vista en gvSIG, en el que se pueda cargar una capa de visualización que contiene el archivo de datos y el mapa sobre el que se realizarán las visualizaciones. En el caso de la implementación provista por defecto, la capa es almacenada en un archivo HDF, que contiene tanto los datos como el modelo digital de elevaciones utilizado como mapa, un *raster* que se muestra de fondo.

Cuando se dispone de un proyecto previamente creado, el primer paso para realizar la visualización lo constituye crear una vista de visualización. La figura 5.5, muestra los pasos para la creación de una vista de visualización coordinada. En el paso uno se debe seleccionar el tipo de documento vista, en el segundo paso crear un nuevo documento vista oprimiendo el botón Nuevo, en el tercer paso, seleccionar el documento creado y luego abrirlo oprimiendo el botón Abrir (paso 4). Realizados estos pasos se muestra una vista preparada para la visualización.

Al estar activa una vista de visualización, se habilita la opción cargar capa. Al seleccionar esta opción se muestra un diálogo para la selección de la capa, que constituye el proyecto de visualización coordinada con el que se desea trabajar. Una vez seleccionado el archivo, se carga el mapa definido y los datos asociados (obsérvese la figura 5.6).

La figura 5.7 muestra los diferentes componentes de una vista luego de cargar una capa en la vista de visualización. A la izquierda se activa el panel de configuración de las visualizaciones. Este panel está separado en tres pestañas correspondientes a las configuraciones generales (1), las configuraciones de las técnicas mostradas sobre el mapa (2) y las configuraciones de las técnicas mostradas como vistas independientes (3). La región principal (4), constituye el área de visualización del mapa y las técnicas proyectadas sobre este. Aquí influyen muchos de los principios presentados en el modelo. En la región principal, se aplica el principio ver todo, pues

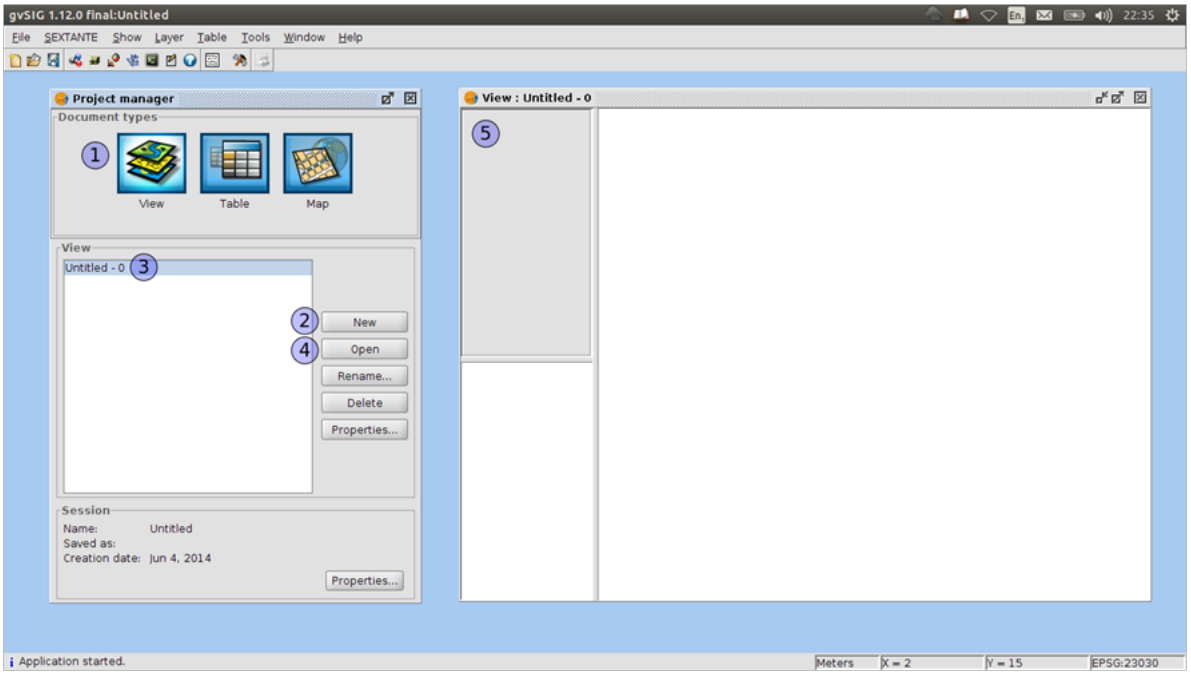

Figura 5.5 Creación de una vista de visualización coordinada.

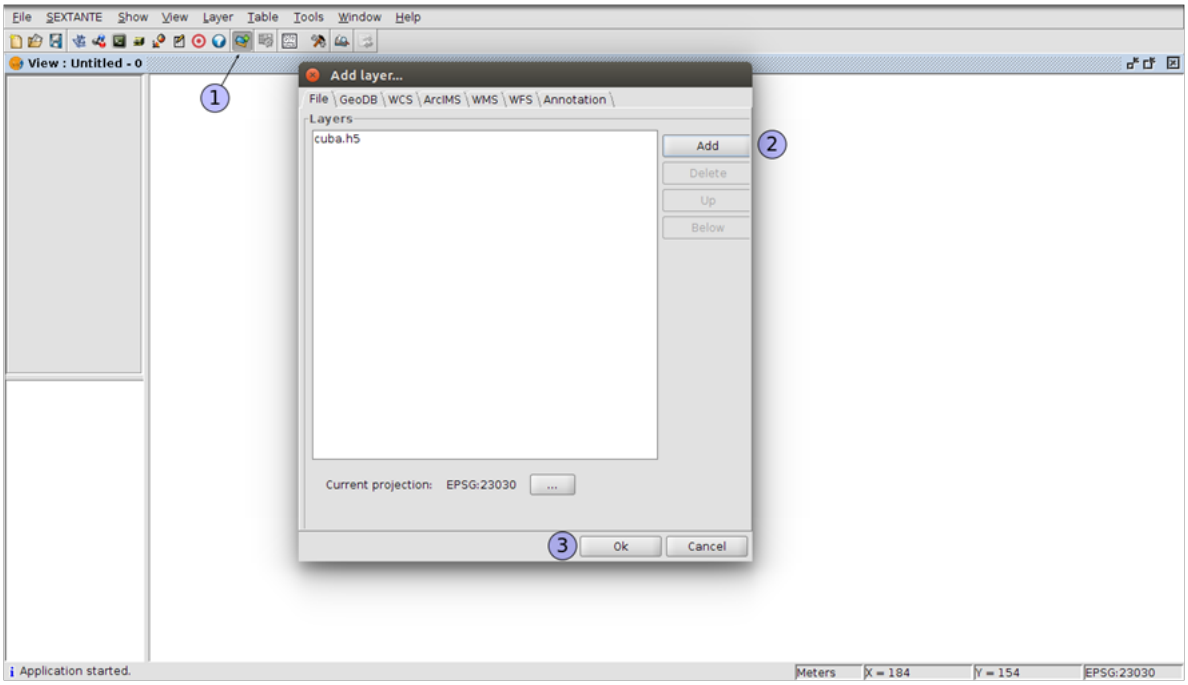

Figura 5.6 Opciones para la adición de un proyecto de visualización coordinada a una vista de visualización.

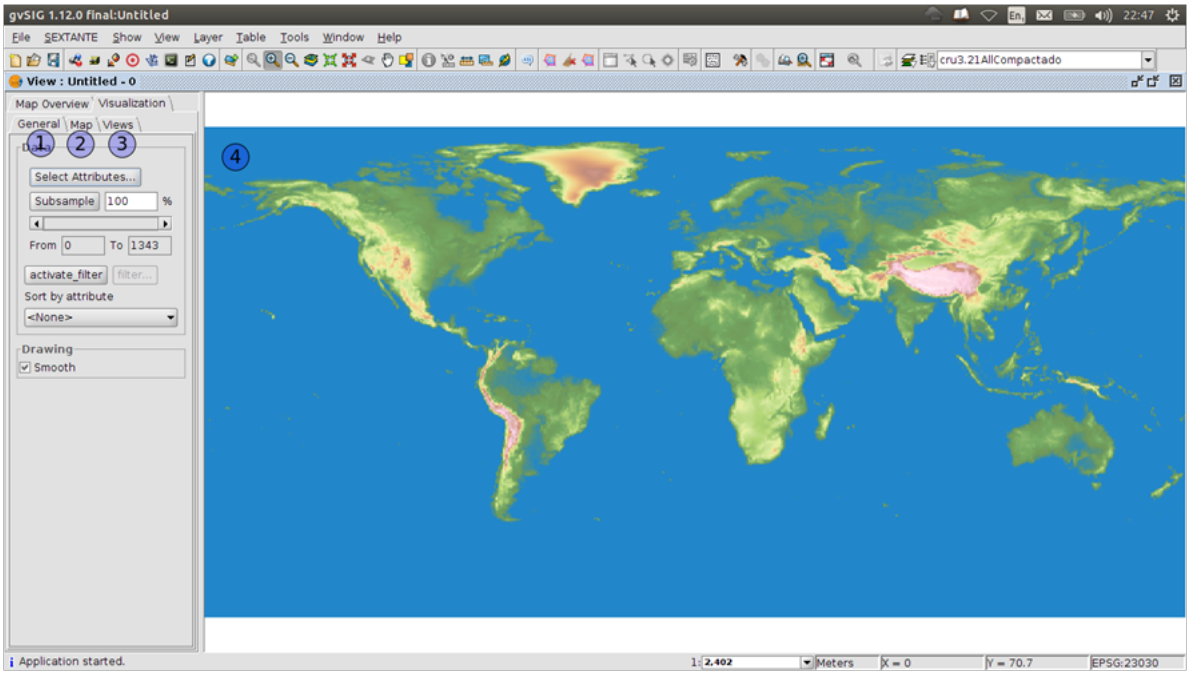

Figura 5.7 Vista de visualización.

posibilita tener una idea de todas las referencias espaciales.

El panel de configuración general (obsérvese la figura 5.8) está dividido en áreas dedicadas a la configuración de la fuente de datos (1) y la configuración de los parámetros de dibujo (2), específicamente el nivel de calidad de los gráficos generados durante el proceso de visualización. Las opciones de configuración de los datos permiten la selección de los atributos a analizar (3), así como establecer el orden entre estos. Adicionalmente, puede obtenerse un subconjunto de los datos (4), al establecer el por ciento de los registros que se quieren visualizar. Otra opción disponible es la aplicación de un filtro (5) sobre los datos a partir de los valores de los distintos atributos.

La aplicación del filtro se realiza a través del cuadro de diálogo de selección de filtros (obsérvese la figura 5.9). En este se permite establecer una expresión booleana a partir de comparaciones entre los atributos y valores suministrados por el usuario. Estas expresiones son a su vez enlazadas por operaciones lógicas de conjunción *AND* o disyunción *OR*.

5.5. Caso de estudio: visualización de grandes volúmenes de datos climáticos mundiales

En este estudio se utiliza el conjunto de datos CRU TS 3.21 (Harris *et al.*, 2014). Este conjunto de datos incluye 1344 registros mensuales de 10 variables climáticas durante 110 años para el período de 1901 hasta 2012. El conjunto de datos cubre la superficie terrestre a una resolución de 0,5 grados. En la tabla 5.2 se muestran dichas variables climáticas. El presente

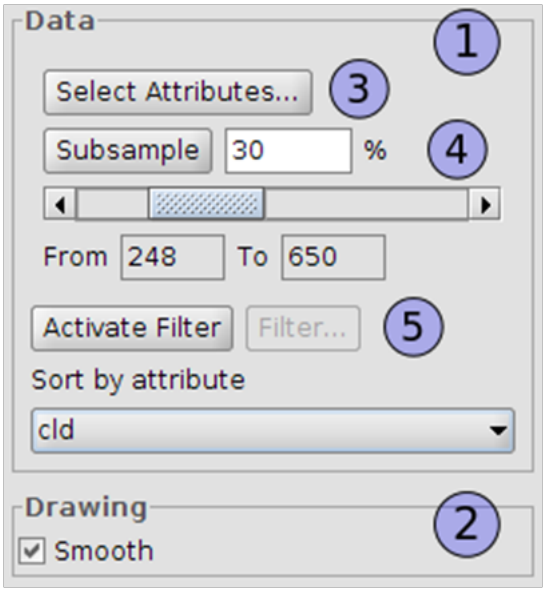

Figura 5.8 Panel de configuración general de las técnicas.

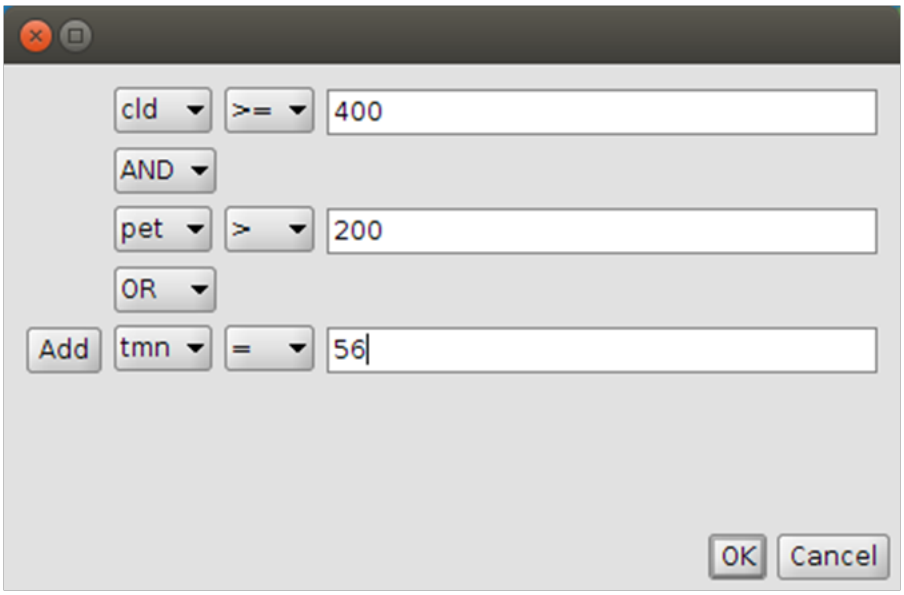

Figura 5.9 Diálogo de filtrado de datos.

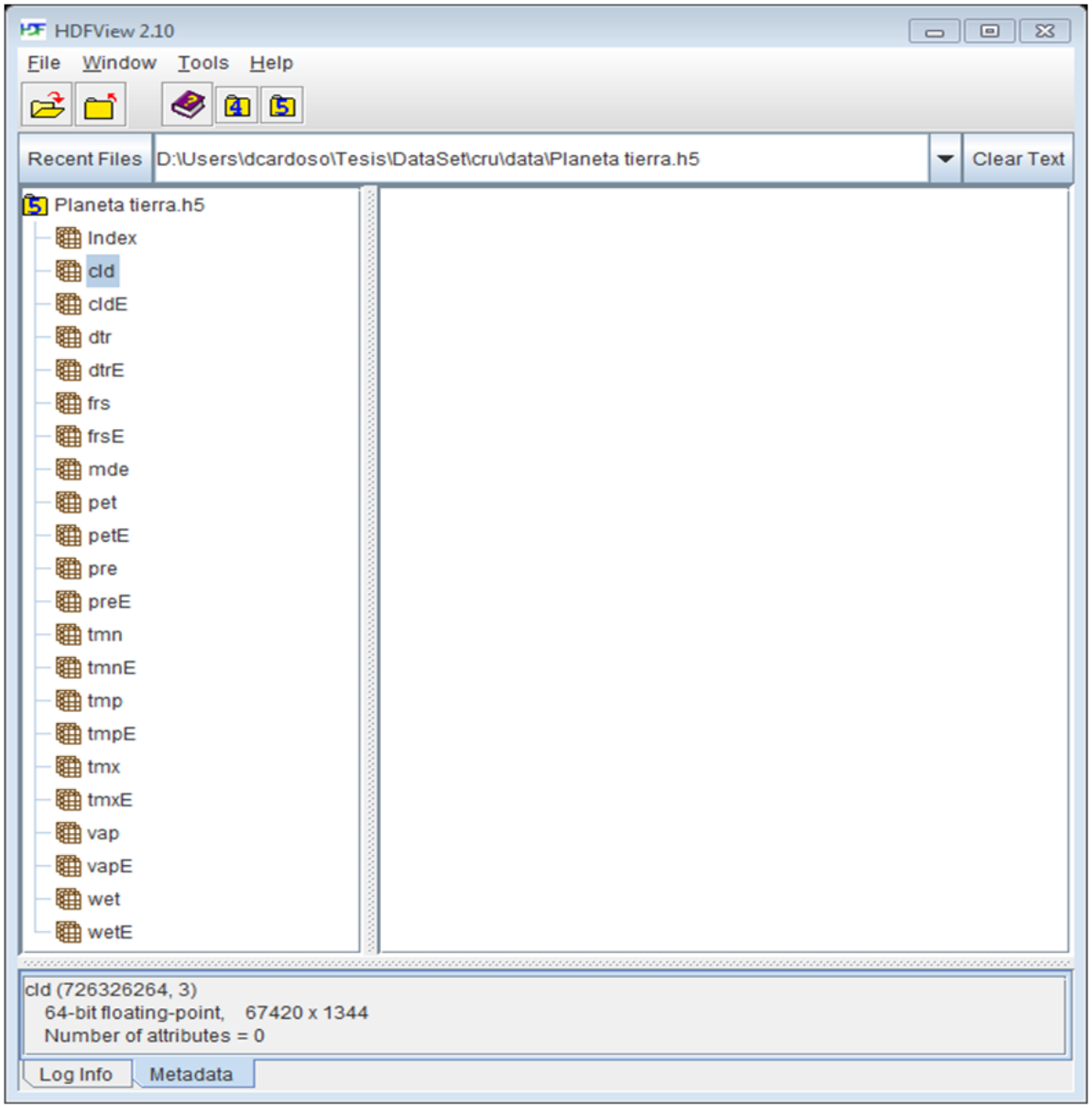

Figura 5.10 HDF creado con los datos climáticos mundiales.

epígrafe tiene como objetivo presentar un caso de estudio donde se realiza el análisis visual de estos datos climáticos mundiales. En cada caso se presentan los componentes y elementos del modelo que se aplican.

Como uno de los resultados de esta investigación, se crearon varios conjuntos de datos HDF. Estos conjuntos de datos almacenan la información con la estructura necesaria para que el módulo de visualización científica de gvSIG, descrito en este capítulo, funcione correctamente. Las herramientas que permiten la transformación de los datos hacia esta estructura son descritas en el siguiente capítulo. La figura 5.10, muestra uno de los conjuntos de datos HDF5 generados con los datos y la estructura necesaria para desarrollar este caso de estudio a escala mundial.

Se crearon diez conjuntos de datos, uno para cada una de las variables. Cada conjunto de datos es una matriz de 67420 filas por 1344 columnas, donde las filas representan a cada una de

las celdas de las que se dispone información y las columnas los 1344 valores mensuales de las variables en cada celda, tomados desde 1901 hasta 2012. Además se crearon diez conjuntos de dados adicionales de 1 por 2 que tienen el valor mínimo y máximo global de cada variable. Estos conjuntos de datos son utilizados para normalizar los valores antes de realizar las visualizaciones. El nombre de estos conjuntos de datos coincide con el nombre de la variable y termina con una letra "E" mayúscula, para representar a los extremos asociados a cada variable.

Además, se crearon dos conjuntos de datos de 360 filas por 720 columnas, nombrados "Index" y "mde". El conjunto de datos "mde", es una matriz que almacena el modelo digital de elevaciones de la región estudiada. Este se utiliza para mostrar como mapa de fondo y sirve de base para seleccionar los datos de las regiones que se desea analizar en las visualizaciones (obsérvese la figura 5.7 (4)). El conjunto de datos "Index", es también una matriz que coincide con las celdas del mapa *raster* utilizado como fondo y que contiene en cada escaque el número de la celda correspondiente, si posee información válida. De esta forma, esta matriz se puede leer, y tener así un acceso directo a los datos de cada una de las celdas.

Los principales componentes del modelo que intervienen en el acceso a estos datos, lo constituyen los formatos de datos científicos y el sistema de información geográfica pues la integración de ambos facilita la gestión de los datos para ser utilizados por las visualizaciones. El módulo para la manipulación de formatos de datos científicos integrado en gvSIG se comunica con el módulo de visualización científica, como muestra la figura 5.1. La herramienta obtenida para la solución a los problemas tratados en este capítulo es utilizada en esta sección para demostrar la viabilidad y utilidad del modelo para el análisis exploratorio de grandes volúmenes de datos.

Validación del sistema. Detectar lo esperado

En este caso de estudio con el objetivo de ejemplificar el uso del sistema, se seleccionaron como técnicas a visualizar matrices de diagramas de dispersión y patrón recursivo. Esta última con el patrón $(12, 1)(1, 112)$ que dispone los datos de los 12 meses de izquierda a derecha y cada año de arriba hacia abajo. Este es un ejemplo donde se aplica el principio establecer estructura.

Para ejemplificar el uso de la herramienta se tomaron cuatro puntos contrastantes, uno sobre el Himalaya, otro sobre el desierto del Sahara, y otros dos sobre la cuenca del Amazonas y Groenlandia respectivamente. Mediante la herramienta que permite modificar el tamaño del gráfico, se aumentaron todos los gráficos para analizar mejor la información de los cuatro puntos en el espacio y el tiempo. De esta forma se cumple con el principio de ampliar y enfocar. Al establecer dos tipos de técnicas diferentes para el análisis se cumple con el principio establecer enlaces. En la figura 5.11 se muestran los resultados de la visualización con patrones recursivos y matrices de diagramas de dispersión.

Si se analizando la visualización de la técnica de patrón recursivo: se puede ver claramente cómo el gráfico sobre el Himalaya y el de Groenlandia poseen valores más bajos durante todo el año para las variables temperatura mínima (tmn), diaria media (tmp), y máxima (tmx), que el gráfico sobre el desierto de Sahara y sobre el Amazonas. Sin embargo, en el gráfico sobre

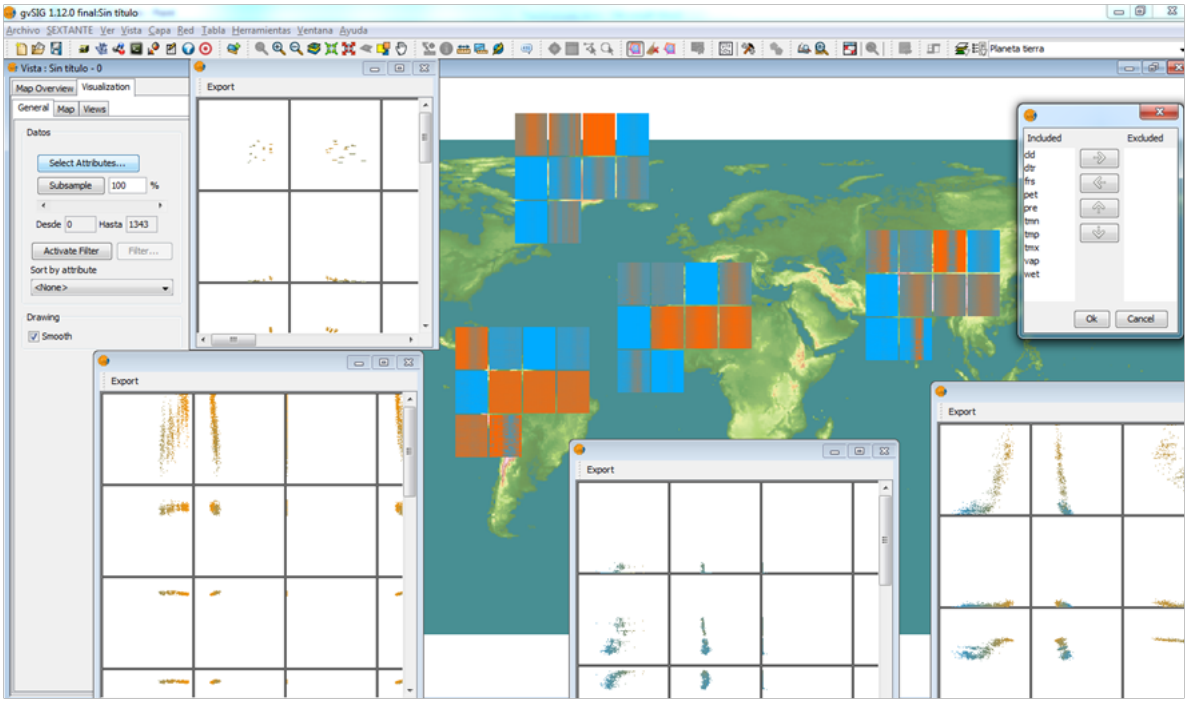

Figura 5.11 Visualización de diferentes puntos utilizando patrón recursivo.

el Amazonas, la humedad (wet) y la presión de vapor de agua (vap) poseen valores más altos de esas variables que todas las demás regiones analizadas. En este análisis se involucran un grupo de tareas sinópticas y tareas del tipo "muestra", pues se aplica el principio de buscar lo reconocible y de ver relaciones.

Se puede observar, además, que el promedio de escarcha diaria (frs), es más intenso en Groenlandia y en el Himalaya que en los otros dos lugares; asimismo, se nota en una disminución considerable de los valores de esta variable en el Himalaya en los meses más calurosos del año para la zona norte del planeta, lo que no es tan pronunciado en Groenlandia. Aquí se puede aplicar el principio atender a particulares resaltando los valores de datos mediante filtros y selecciones que refinen más el análisis.

El recuadro de la parte superior derecha de la figura 5.12, muestra el orden de las variables, que en patrones recursivos se distribuyen de izquierda a derecha y de arriba hacia abajo. Este orden se puede variar y reordenar lo que da la posibilidad de establecer estructura para comparar variables relacionadas y facilitar el análisis de estas.

Se puede observar claramente que hay una correspondencia entre los valores de temperatura (media, mínima o máxima) y el promedio de escarcha diario. Por ejemplo, se observa que a medida que aumenta la temperatura por el eje x (de izquierda a derecha), disminuye la escarcha por el eje y. Esto constituye un resultado esperado, ya que a menor temperatura mayor será el nivel de congelación. Este es un ejemplo más de los principios ver relaciones y buscar lo reconocible.

Validación del sistema. Descubrir lo inesperado

En esta sección se analiza un nuevo caso de estudio centrado en Cuba. Para la realización del análisis fueron seleccionadas las técnicas de patrones recursivos, *table lens* y espiral de tiempo.

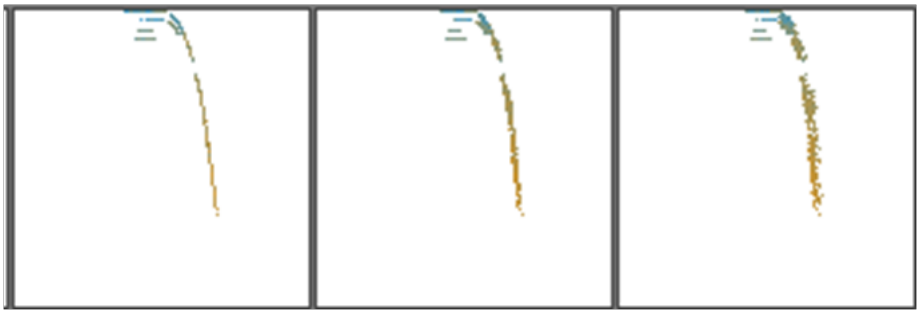

Figura 5.12 Gráfico de matrices de diagrama de dispersión asociado al Himalaya, las 3 variables de temperaturas correlacionadas con la variable de escarcha diaria.

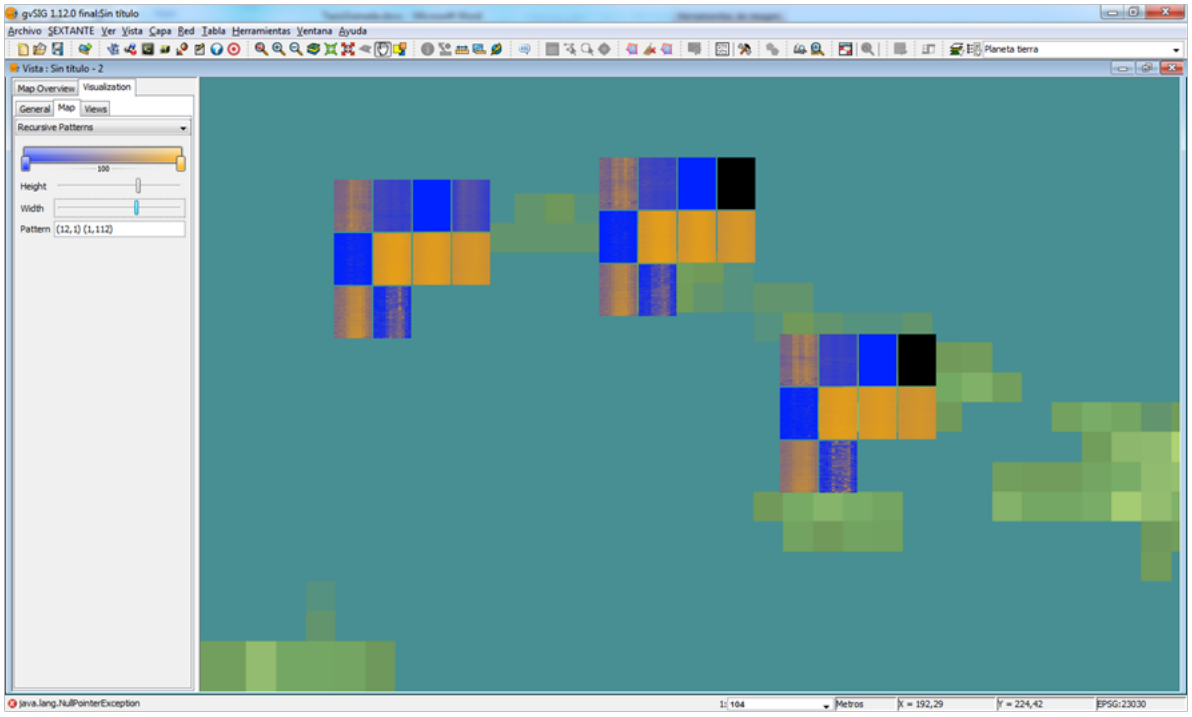

Figura 5.13 Visualización de diferentes puntos del territorio cubano utilizando patrón recursivo.

A la técnica de patrones recursivos se le aplicó el patrón $(12, 1)(1, 112)$ para la disposición de los datos de los 12 meses de izquierda a derecha y los 112 años de arriba hacia abajo. La técnica de visualización espiral de tiempo fue configurada con un período de 12. De esta manera se hace corresponder cada vuelta de la espiral con uno de los años analizados. Se escogieron tres puntos representativos del territorio cubano, uno en oriente (zona montañosa), otro en el centro (llanura) y el último en occidente (llanura). El resultado se muestra en la figura 5.13.

Se aplicaron las técnicas de visualización sobre distintos puntos de la geografía cubana en la búsqueda de patrones o fenómenos "interesantes". En diversos puntos de la geografía oriental fue detectado un comportamiento peculiar. Al realizar el análisis con la técnica de visualización de patrones recursivos pudo observarse un valor significativo en la variable correspondiente a las precipitaciones. El valor está presente en varios puntos en el mismo espacio de tiempo, lo que indicaría un fenómeno climático que afectó gran parte de la región mencionada. En la figura 5.14 se muestra la visualización en uno de los puntos analizados; la coloración amarilla del

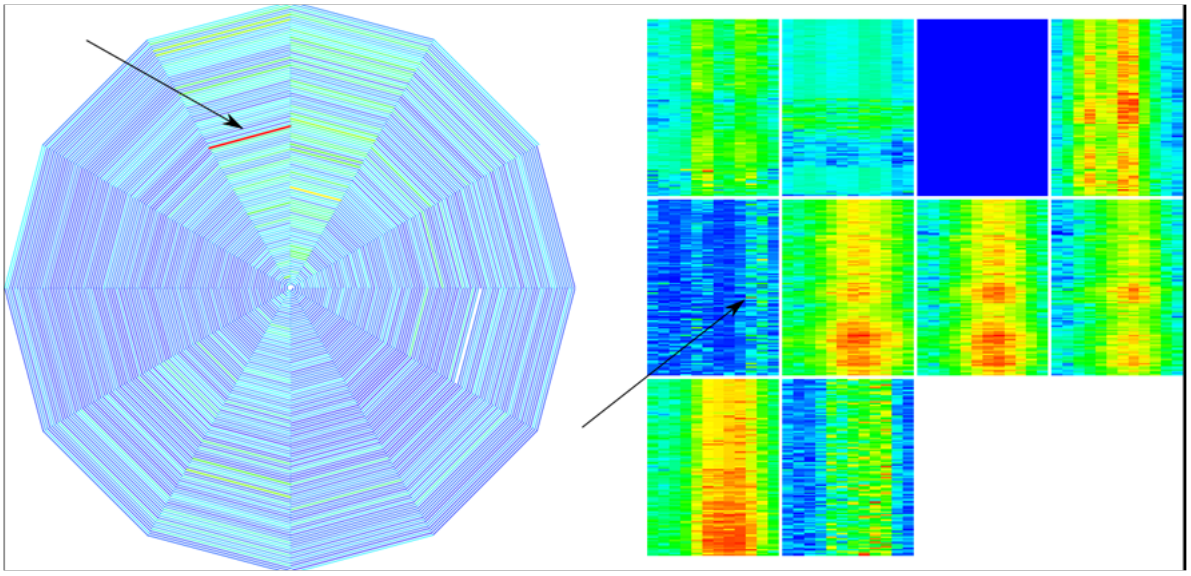

Figura 5.14 Visualización de registros climáticos de la región oriental de Cuba utilizando patrones recursivos y espiral de tiempo.

registro mencionado indica un valor inusualmente alto. En este caso se está cumpliendo con el principio ampliar y enfocar y atender a particulares.

La realización del análisis de la misma región con la técnica de visualización *table lens* (obsérvese la figura 5.15), permitió constatar que corresponde al mes de octubre de 1963 -registro 753. Para el punto mostrado este mes registró un acumulado de precipitaciones de $747,9$ mm y coincide con el paso del huracán Flora por la región oriental del país. Aquí se demuestra nuevamente un ejemplo del principio atender a particulares. En este caso se ha utilizado una tarea elemental pero que ha servido para obtener información específica de elementos de datos.

Como se puede observar mediante la técnica de espiral de tiempo (obsérvese figura 5.14) el valor es comparativamente superior a otros meses lluviosos del mismo año y destaca con respecto a registros del mismo mes durante los años analizados. Este valor inusual marca un indicio para considerar que en ese momento ocurrió un fenómeno meteorológico de interés regional. Al consultar los datos históricos se corroboró que coincidía con el huracán que pasó y estuvo estacionario en esa región en octubre de 1963.

5.6. Conclusiones parciales

En este capítulo se brinda una solución a los problemas del análisis exploratorio de datos con alta densidad espacial. Para eso se aplicó el modelo propuesto en el capítulo 3 y se obtuvo una herramienta que permite la extracción de conocimiento, reconocimiento de patrones, tendencias, anomalías y relaciones entre múltiples variables distribuidas uniformemente sobre un territorio; estas variables pueden ser series temporales y por la alta densidad espacial de que se dispone, los volúmenes de información con los que se trabaja pueden llegar a ser difíciles de manipular.

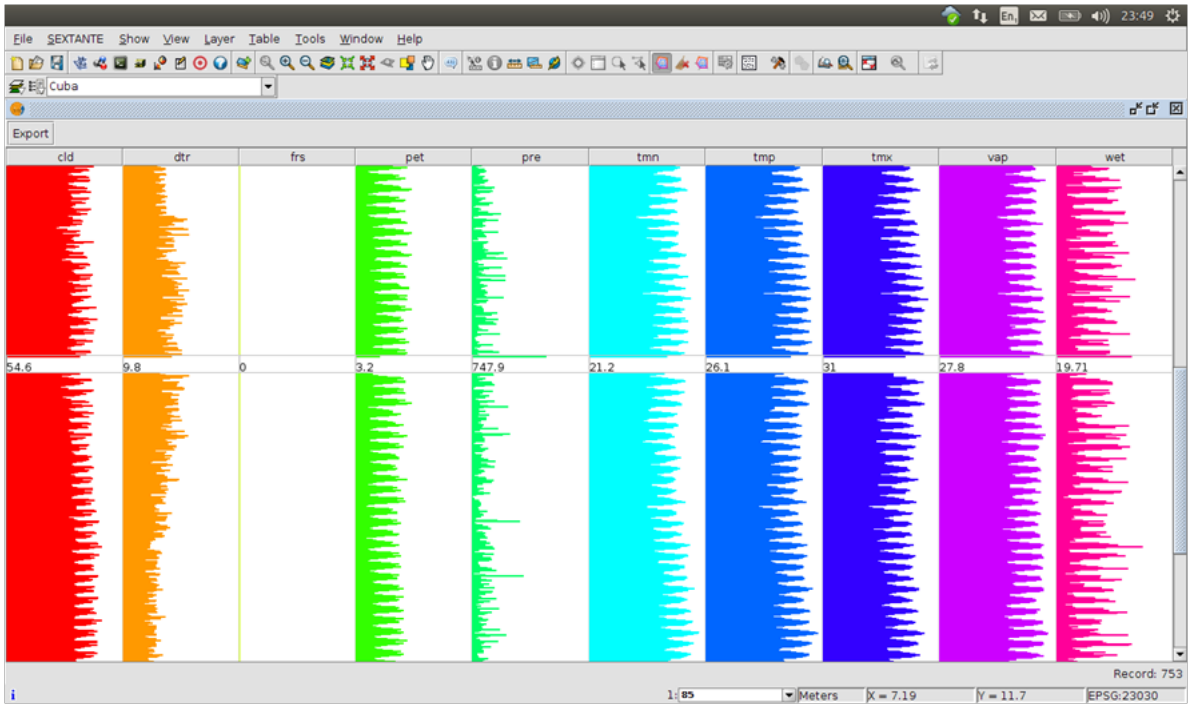

Figura 5.15 Visualización de registros climáticos de la región oriental de Cuba utilizando *table lens*. El foco se encuentra aplicado al registro 753 (octubre de 1963).

Se propone la integración de un conjunto de técnicas de visualización de datos multiparamétricos en un sistema de información geográfica libre. El sistema de información geográfica seleccionado para esta integración fue la versión 1,12 de gvSIG. Las herramientas fueron desarrolladas como una extensión de este sistema que permite realizar análisis exploratorio de grandes volúmenes de datos espacio-temporales. Estos grandes volúmenes de datos son gestionados mediante la incorporación de herramientas para la manipulación de formatos de datos científicos. En particular desde el módulo de visualización científica de gvSIG desarrollado para este caso, se puede acceder a los grandes volúmenes de datos mediante un conjunto de herramientas que se implementaron para manipular el formato HDF. En este capítulo se describe, además, el formato HDF propio creado para ser analizado mediante esta extensión. Las herramientas implementadas facilitan que los usuarios analistas puedan formular y/o corroborar hipótesis relacionadas con los datos analizados.

La efectividad de esta herramienta se ha podido demostrar a través de un caso de estudio que evidencia su utilidad para realizar análisis exploratorio de grandes volúmenes de datos, donde se puede detectar patrones esperados o descubrir relaciones desconocidas que llaman la atención de los usuarios especialistas de un dominio de aplicación. La herramienta obtenida se ha descrito haciendo referencias a los principales componentes del modelo propuesto, y constituye una forma de implementación del modelo para solucionar el problema planteado en este caso.

6 Herramientas para el soporte de archivos de formatos de datos científicos en sistemas de información geográfica

En el capítulo anterior se hizo uso de un grupo de funciones para la manipulación de formatos de datos científicos. En particular, se utilizó el formato de datos HDF para el almacenamiento y manipulación de los grandes volúmenes de datos que intervienen en el análisis exploratorio de datos con alta densidad espacial. En este capítulo se describe un grupo de herramientas que son utilizadas para la preparación y gestión de los datos que se requieren en un módulo de visualización científica de un sistema de información geográfica como el descrito en el capítulo anterior.

De los formatos de datos científicos espacio-temporales analizados en el epígrafe 2.7, se seleccionaron HDF y NetCDF como formatos para la incorporación en sistemas de información geográfica. Las principales razones que influyeron en esta selección fueron: disponer de una amplia biblioteca implementada en Java para acceder a ambos formatos; tener conocimiento previo del uso e integración con éxito de estos formatos de datos científicos con sistemas de información geográfica comerciales; ser superiores que los otros formatos estudiados en cuanto a las facilidades que poseen para el almacenamiento de grandes volúmenes de datos.

De los sistemas de información geográfica estudiados en el epígrafe 2.4, la alternativa más viable para la integración de los formatos HDF y NetCDF lo constituye gvSIG y la biblioteca Sextante, debido a la perfecta integración que existe entre ellos. Las herramientas desarrolladas en este capítulo fueron implementadas como una extensión de la biblioteca Sextante. La ventaja de que tanto gvSIG como Sextante estén implementados en Java, se facilita su integración con los formatos de datos científicos HDF y NetCDF. Esto se puede lograr mediante las bibliotecas HDF-Java y NetCDF 4.2. Otra ventaja adicional de la incorporación de estos nuevos algoritmos en Sextante, está dada por la posibilidad de incorporar esta biblioteca en otros sistemas de información geográfica libres, basados en Java, que se acoplen con Sextante.

6.1. Selección de las tecnologías

Las herramientas propuestas en este capítulo se pueden agrupar según el tipo de los objetos de entrada. En este epígrafe se muestran los principales diagramas de actividades para el trabajo con tablas, archivos *raster*, archivos HDF y archivos NetCDF.

Trabajo con tablas

La figura 6.1 muestra cómo se realiza el trabajo cuando la entrada es un objeto tipo tabla, el primer paso es seleccionar una tabla que se haya cargado previamente o simplemente seleccionarla del disco duro o de una conexión a una base de datos. El siguiente paso puede ser calcular estadísticos sobre la tabla correspondiente o pasar a la conversión de la tabla en un conjunto de datos dentro de un formato de dato científico (SDF por sus siglas en inglés). En este momento se selecciona dónde se desea guardar el archivo resultante, se introducen todos los parámetros necesarios y se ejecuta el algoritmo.

Trabajo con archivos de tipo *raster*

El diagrama de actividades para el trabajo con archivos *raster*, es prácticamente igual al diagrama anterior, solamente difiere en que el archivo que se va a seleccionar como entrada es un *raster* que ya ha sido cargado en el sistema, o que se especifica su camino en el sistema de archivos.

Trabajo con archivos NetCDF

Cuando la fuente de entrada es un archivo NetCDF, el primer paso es seleccionar el archivo NetCDF con el que se va a trabajar, luego se debe seleccionar el conjunto de datos que se desea convertir. En este paso se debe elegir una de cuatro opciones posibles. Como se muestra en la figura 6.2 una de las posibilidades es convertir el conjunto de datos seleccionado a una tabla. Otra posibilidad es convertirlo a un *raster*. Además existe la opción de convertirlo a un HDF estándar (con la misma forma que tiene el conjunto de datos original en el formato NetCDF) o convertirlo a un HDFcru con la estructura permitida para ser analizado posteriormente con el módulo de visualización científica. Luego de elegir cualesquiera de estas cuatro posibilidades se deben introducir los parámetros necesarios para el algoritmo (como el nivel de compresión si se va a convertir hacia un HDF, dirección donde va a almacenar el archivo, etc.) y posteriormente se procede con la ejecución del algoritmo de conversión.

Trabajo con archivos HDF

Cuando la fuente de datos de entrada es un archivo HDF, el proceso es similar al del caso anterior (obsérvese la figura 6.3). El primer paso es seleccionar el archivo de entrada y luego el conjunto de datos que se desea convertir. En este punto el conjunto de datos se puede convertir a una tabla, a un *raster*, o en un NetCDF (con la misma estructura que el conjunto de dato que posee el archivo en el formato HDF original). Además, si el HDF de entrada posee la estructura necesaria para ser analizado por el módulo de visualización científica, existe la posibilidad de hacer un recorte de este, para almacenar en el archivo de salida los datos para una región de interés. En este caso el resultado será un archivo HDF con la estructura necesaria para ser

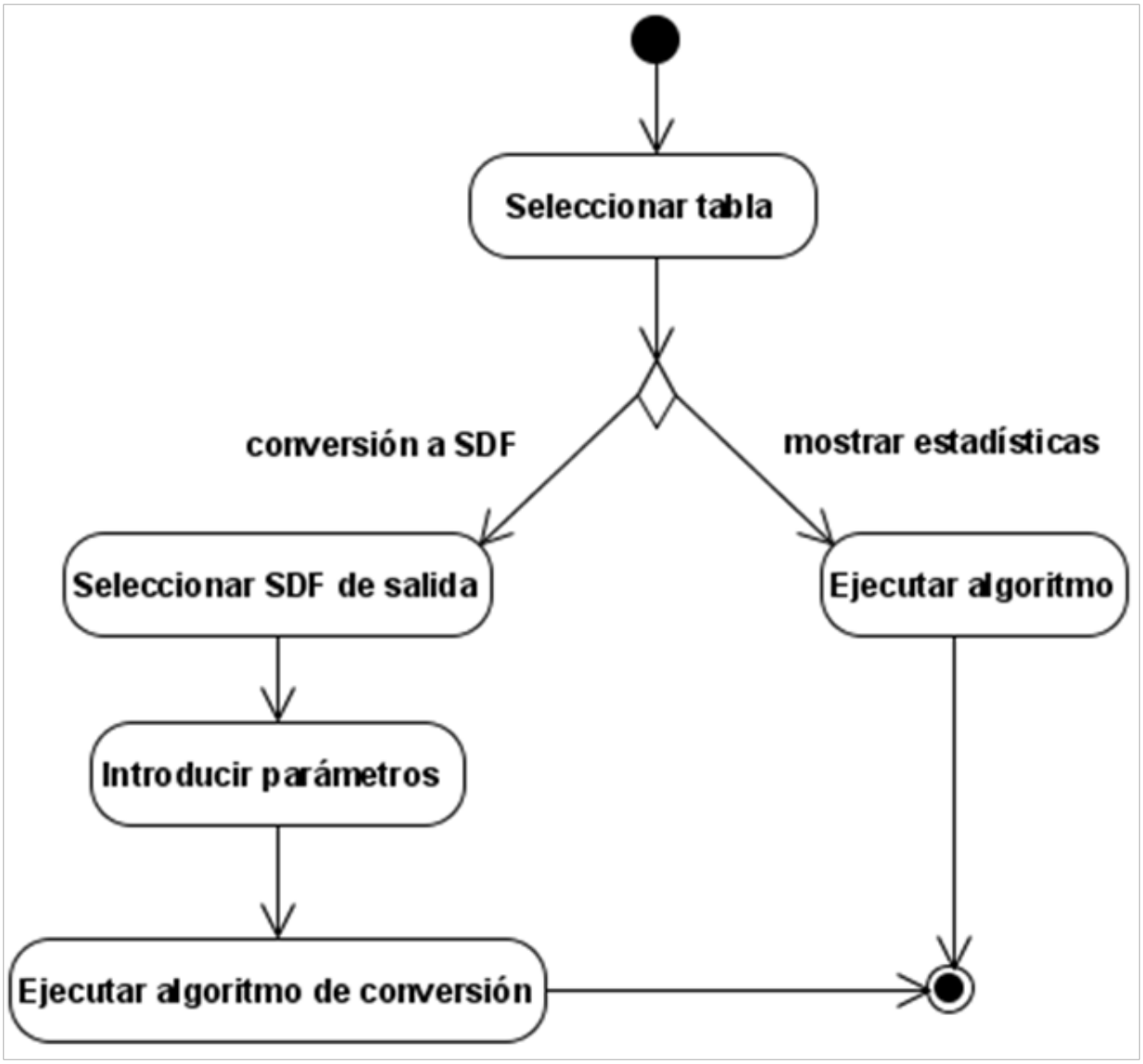

Figura 6.1 Diagrama de actividad para trabajo con tablas. Para referirse a los formatos de datos científicos en el gráfico se han utilizado las siglas SDF del término en inglés *Scientific Data Format*.

utilizado en el módulo de visualización científica. Luego de decidir la opción que se desea realizar, se introducen los parámetros y se ejecuta el algoritmo de conversión.

6.2. Descripción de los algoritmos

En este epígrafe se describen los principales algoritmos implementados como una extensión de Sextante. El listado de la figura 6.4 muestra el algoritmo para transformar de un formato de dato científico de entrada para otro formato de dato científico. Los parámetros de entrada son las direcciones de los archivos de entrada y de salida y el nivel de compresión que se desea aplicar. En la función *SDFin-To-SDFout*, lo primero que se hace es obtener el tipo de dato almacenado en el conjunto de datos. Luego se obtienen las dimensiones del formato de dato de

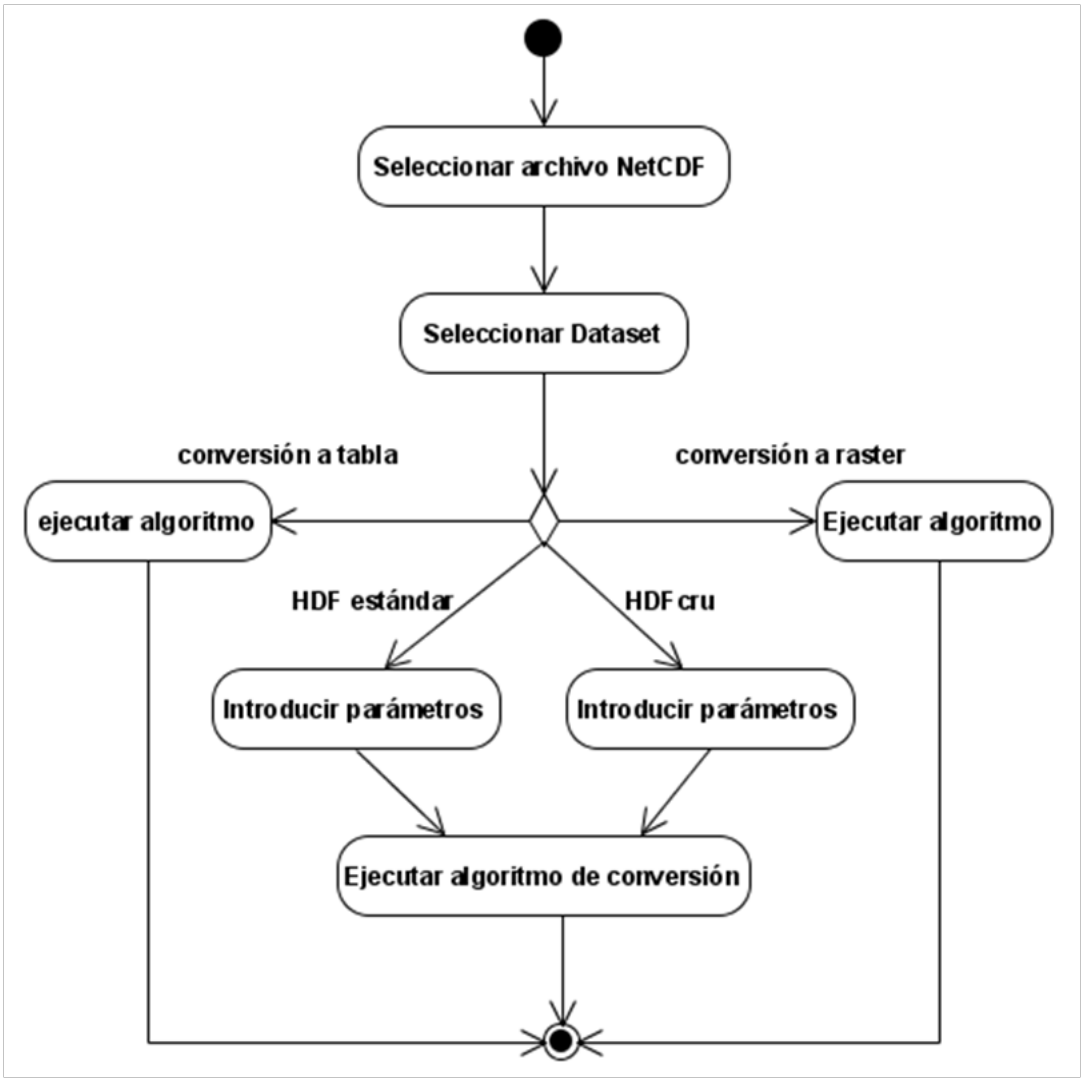

Figura 6.2 Diagrama de actividades para el trabajo con archivos NetCDF. Para referirse a los formatos de datos científicos en el gráfico se han utilizado las siglas SDF del término en inglés *Scientific Data Format*.

entrada. Posteriormente se crea en el archivo de salida un conjunto de datos con el mismo tipo y dimensiones que el formato de entrada. Se especifica además el nivel de compresión con que se desea almacenar los datos en el archivo de salida. Se obtienen el número de dimensiones y se hace una lectura y escritura por bloques sobre los archivos correspondientes.

El listado de la figura 6.5, muestra el algoritmo para convertir archivos *raster* a un formato de dato científico. Las entradas del algoritmo son una lista de archivos *raster*, el nivel de compresión con que se desea guardar la información y el camino donde se desea guardar el archivo del formato de dato científico de salida. Para cada *raster* de la lista de entrada, se obtiene el tipo de datos que almacena y el número de bandas. Si el archivo *raster* posee una sola banda, el algoritmo tendrá un tratamiento diferenciado con respecto a otros *raster* que posean múltiples bandas. En el caso de una sola banda, se obtienen las dimensiones X e Y, y se manda a crear un conjunto de datos en el formato de dato científico de salida con el tipo de dato, las dimensiones

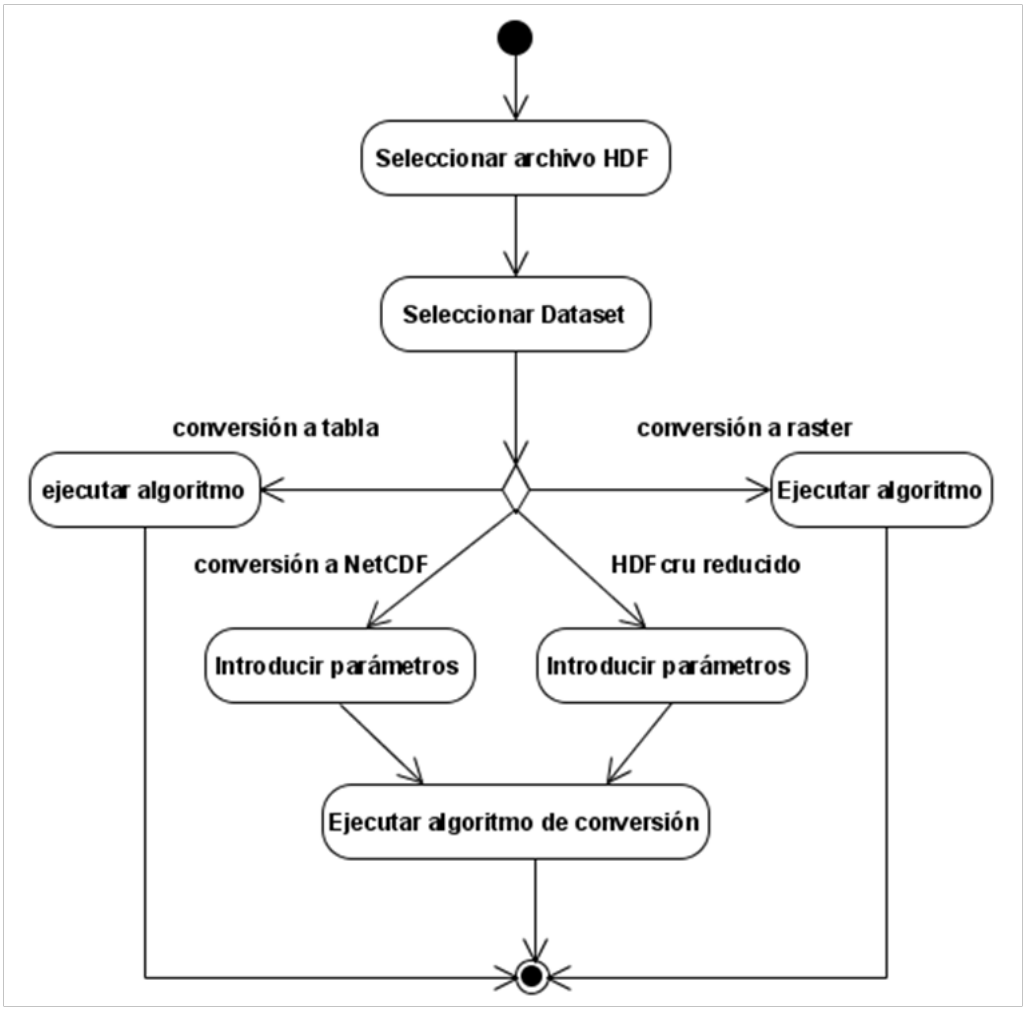

Figura 6.3 Diagrama de actividades para el trabajo con archivos HDF. Para referirse a los formatos de datos científicos en el gráfico se han utilizado las siglas SDF del término en inglés *Scientific Data Format*.

fIN: Archivo de entrada.
$fOUT$: Archivo de salida.
n: Nivel de compresión.

SDFIN-TO-SDFOUT$(fIN, fOUT, n)$
1 $t \leftarrow$ DATA-TYPE(fIN)
2 $dim \leftarrow$ GET-DIMENSION(fIN)
3 CREATE-DATASET-VARIABLE$(fOUT, t, dim, n)$
4 $numdim \leftarrow$ SIZE(dim)
5 READ-BLOCK$(numdim, fIN)$
6 WRITE-BLOCK$(numdim, fOUT)$

Figura 6.4 Algoritmo para transformar de un formato de dato científico de entrada para un formato de salida.

L: Lista de *rasters*
n: Nivel de compresión
f: Archivo de salida.

```
RASTER-TO-SDF(L, n, f)
 1  for i ← 1 to SIZE(L)
 2       do r ← L[i]
 3          t ← DATA-TYPE(r)
 4          nbands ← NUMBER-OF-BANDS(r)
 5          NX ← NX(r)
 6          NY ← NY(r)
 7          if nbands = 1
 8            then dim ← {NX, NY}
 9                 CREATE-DATASET(f, t, dim, n)
10                 for j ← 1 to NX
11                     do for h ← 1 to NY
12                              do line[h] ← CELL-VALUE(j, h)
13                        WRITE-LINE(f, x, line)
14            else
15                 dim ← {NX, NY, nbands}
16                 CREATE-DATASET(f, t, dim, n)
17                 for z ← 1 to nbands
18                     do for j ← 1 to NX
19                              do for h ← 1 to NY
20                                       do line[h] ← CELL-VALUE(j, h, z)
21                                 WRITE-LINE(f, x, z, line)
```

Figura 6.5 Algoritmo para convertir un *raster* para un formato de dato científico.

y el nivel de compresión deseado. Para cada dimensión se obtiene el valor de celda y se escribe una línea de datos para el fichero de salida. En caso de que exista más de una banda, el proceso es similar, pero se necesitan tres ciclos anidados para escribir los datos en un conjunto de datos de tres dimensiones en el fichero de salida.

El algoritmo que permite la transformación de datos NetCDF con la estructura original de los datos que suministra CRU, para un archivo HDF con el formato necesario por el módulo de visualización científica se presenta en el listado de la figura 6.6. Las entradas son: el archivo de entrada NetCDF con la estructura del CRU, el archivo de salida HDF y el nivel de compresión con que se desean almacenar los datos. En un primer paso se obtiene el tipo de dato de los conjuntos de datos, los puntos con información válida, es decir las celdas que corresponden con la superficie terrestre del planeta, donde se ha recopilado información; estos conjuntos de datos no poseen información en los océanos.

Se capturan los momentos de tiempo y las dimensiones espaciales. Luego se manda a crear un conjunto de datos en el HDF de salida con el tipo de dato, las dimensiones y el nivel de compresión. Se obtienen los extremos globales de cada variable y luego se pasa a un ciclo que

```
NETCDFCRU-TO-HDF(fIN, fOUT, n)
 1  t ← DATA-TYPE(fIN)
 2  L ← INFORMATION-POINTS(fIN)
 3  time ← TIME(fIN)
 4  dim ← {SIZE(L), time}
 5  CREATE-DATASET-VARIABLE(fOUT, t, dim, n)
 6  minmax ← {MaxValue, MinValue}
 7  for i ← 1 to SIZE(L)
 8       do point ← L[i]
 9          line ← READ-TIME(fIN, point)
10          WRITE-LINE(fOUT, i, line)
11          for j ← 1 to SIZE(line)
12               do minmax[0] ← MIN(minmax[0], line[j])
13                  minmax[1] ← MAX(minmax[1], line[j])
14  dim ← {2}
15  CREATE-DATASET-EXTREME(fOUT, t, dim, n)
16  WRITE-EXTREME(fOUT, minmax)
17  if not CONTAINS-INDEX(fOUT)
18     then idx ← GET-INDEX(L)
19          NX ← NX(fIN)
20          NY ← NY(fOUT)
21          dim ← {NY, NX}
22          CREATE-DATASET-INDEX(fOUT, t, dim, n)
23          WRITE-INDEX(fOUT, idex)
```

Figura 6.6 Algoritmo para transformar datos de NetCDF con la estructura original de los datos que suministra CRU, para un archivo HDF con el formato necesario por el módulo de visualización científica.

lee línea por línea del archivo NetCDF y esta información se escribe con la estructura necesaria en el archivo HDF de salida. En cada paso se van actualizando los extremos de cada variable. En el archivo de salida, se crea un conjunto de datos con los valores extremos de cada variable y posteriormente se almacenan los valores calculados. Si el archivo HDF de salida no contiene el conjunto de datos "Index", se manda a crear el índice y se guarda como otro conjunto de datos.

Principales funcionalidades del módulo

El módulo realizado cuenta con 19 algoritmos, en la tabla 6.1 aparece cada uno de ellos junto con una pequeña descripción de la función que realizan.

Basic StatisticsRaster

Basic StatisticsRaster, es un algoritmo que inicialmente cuando se abre la caja de herramienta no puede ser ejecutado, a menos que se haya cargado con antelación una capa *raster* en gvSIG, o se haya ejecutado antes otro algoritmo que como resultado genere un *raster*. El mismo tiene dos parámetros: uno de entrada que es el *raster* a seleccionar y otro de salida, que es la dirección donde se desea guardar la tabla resultante con todos los estadísticos calculados referentes al

Tabla 6.1 Principales funcionalidades del módulo de manipulación de formatos de datos científicos.

Nombre	**Función**
Basic StatisticsRaster	Muestra estadísticas del *raster*
Basic StatisticsTable	Muestra estadísticas de una tabla
ClipcruHDF	Recorta el HDF a una región de interés
HDF toRaster	Convierte un HDF a un *raster*
HDF toTable	Convierte un HDF a una tabla
NetCDFcruto HDF	Convierte un NetCDF generado por el CRU a un HDF con la estructura permitida por el módulo de visualización científica
NetCDFto HDF	Convierte un NetCDF a un HDF
NetCDFtoRaster	Convierte un NetCDF a un *raster*
NetCDFtoTable	Convierte un NetCDF a una tabla
Rastersto HDF	Convierte múltiples *rasters* a un HDF
RasterstoNetCDF	Convierte múltiples *rasters* a un NetCDF
Rasterto HDF	Convierte un *raster* a un HDF
RastertoNetCDF	Convierte un *raster* a un NetCDF
RotateRaster	Rota un *raster* 90 grados en contra de las manecillas del reloj
Tablesto HDF	Convierte múltiples tablas a un HDF
TablestoNetCDF	Convierte múltiples tablas a un NetCDF
Tableto HDF	Convierte una tabla a un HDF
TabletoNetCDF	Convierte una tabla a un NetCDF

raster.

Basic StatisticsTable

Basic StatisticsTable es un algoritmo similar al anterior, necesita como entrada una tabla en gvSIG. El algoritmo tiene dos parámetros: uno de entrada que es la tabla a seleccionar y otro de salida que es la dirección donde se desea guardar la tabla resultante con todos los estadísticos calculados sobre los valores de la tabla de entrada.

ClipcruHDF

El algoritmo ClipcruHDF sí puede ser ejecutado desde que se abre la caja de herramientas, puesto que no exige ningún objeto de entrada (*raster* o tabla). Necesita como parámetro de entrada la dirección de un archivo HDF con la estructura permitida por el módulo de visualización científica, las coordenadas del punto superior izquierdo y el punto inferior derecho de la región de interés, que se necesita extraer del HDF de entrada. Como parámetro de salida este requiere la dirección donde se guardará el nuevo archivo generado. Para obtener mejores resultados, este algoritmo debe usarse en combinación con el módulo de visualización científica, puesto que este proporciona un mecanismo que devuelve las coordenadas de los puntos requeridos, cuando se selecciona un área específica sobre el modelo digital de elevaciones (MDE).

HDFtoNetCDF

El algoritmo HDFtoNetCDF, también puede ser ejecutado desde que se inicia Sextante. En este caso, los parámetros de entrada son: la dirección del archivo HDF que se desea convertir y el *dataset* dentro de este archivo. Como parámetros de salida, se tiene la dirección donde se guardará el NetCDF creado por el algoritmo. En este algoritmo el archivo de salida tendrá un *dataset* con las mismas características del *dataset* de entrada.

HDFtoRaster

El algoritmo HDFtoRaster tiene iguales parámetros de entrada que el algoritmo anterior, como salida este debe recibir la dirección donde se guardará el *raster* resultante; de no especificarse este parámetro se creará de manera temporal. El *raster* puede verse luego de terminar la ejecución del algoritmo en una vista de gvSIG.

HDFtoTable

HDFtoTable es igual al anterior, solo que se especifica la dirección donde se guardará la dirección de la tabla. Esta tabla de salida se puede examinar en gvSIG, una vez terminada la ejecución del algoritmo.

NetCDFcrutoHDF

El algoritmo NetCDFcrutoHDF fue explicado anteriormente; como característica singular este debe recibir como parámetro de entrada un NetCDF generado por *Climatic Research Unit* (CRU). Como parámetro de salida tendrá la dirección del archivo HDF creado con la estructura del módulo de visualización científica, se puede especificar si se desea generar el índice o no y el nivel de compresión del archivo.

NetCDFtoHDF

El algoritmo NetCDFtoHDF fue diseñado para transformar datos de un archivo NetCDF para uno HDF manteniendo la estructura original del NetCDF. Los parámetros de entrada son la dirección del archivo NetCDF y el nombre de un conjunto de datos dentro del mismo archivo, que es el que se desea convertir. Como parámetros de salida se tiene la dirección del HDF donde se creará el nuevo conjunto de datos y el nivel de compresión.

NetCDFtoRaster

NetCDFtoRaster convierte un conjunto de datos de un archivo NetCDF en un *raster*. Los parámetros de entrada del algoritmo son la dirección del archivo NetCDF y como salida se especifica la dirección donde se guardará el *raster* resultante; de no especificarse, este se creará de manera temporal. El *raster* puede verse luego de terminar la ejecución del algoritmo en una vista de gvSIG.

NetCDFtoTable

NetCDFtoTable es similar al algoritmo anterior, solo difiere en que recibe la dirección donde se guardará la tabla de salida en lugar de un *raster*.

RastertoHDF y RasterstoHDF

Estos dos algoritmos son prácticamente iguales, el primero recibe como entrada un *raster* y el segundo múltiples *rasters*. Para ser ejecutados ambos necesitan que previamente se haya cargado o generado por otro algoritmo, al menos una capa *raster*. Como parámetros de salida,

requieren la dirección donde se guardará el archivo HDF resultante que contendrá el conjunto de datos creado, así como el nivel de compresión.

RastertoNetCDF y RasterstoNetCDF

Con estos dos algoritmos, se pueden pasar uno o varios archivos *raster*, para un archivo NetCDF; en este caso no se contempla la posibilidad de comprimir los datos en el formato NetCDF.

TabletoHDF y TablestoHDF

Estos algoritmos permiten copiar una o múltiples tablas para un archivo HDF. Ambos para ser ejecutados, necesitan que previamente se haya cargado o generado por otro algoritmo al menos una tabla. Como parámetros de salida se tiene la dirección donde se guardará el archivo HDF resultante, que contendrá el conjunto de datos creado. Estos algoritmos admiten comprimir los datos con un nivel de compresión suministrado por el usuario.

TabletoNetCDF y TablestoNetCDF

Estos dos algoritmos son similares a los anteriores, difieren en el formato de salida, que es un archivo NetCDF que no permite nivel de compresión.

En este epígrafe se han descrito los principales algoritmos implementados como una extensión de Sextante. Con el objetivo de justificar la utilidad de las herramientas desarrolladas, se propone en el siguiente epígrafe realizar un caso de estudio para la preparación de un conjunto de datos sobre la península Ibérica, que pueda ser utilizado por el módulo de visualización científica de gvSIG.

6.3. Caso de estudio: creación de un conjunto de datos para el análisis exploratorio de datos climáticos de España con alta densidad espacial

La figura 6.7, muestra la estructura de la variable “cld” almacenada en un archivo NetCDF con el formato original que brinda la unidad de investigaciones climática (CRU por sus siglas en inglés). Se dispone de diez variables climáticas con estructura similar, las cuales están almacenadas por separado, cada una en un archivo NetCDF. Para mostrar este archivo se ha utilizado el programa HDFView, en el cual se tiene abierto el conjunto de datos que almacena los datos de la variable “cld”. Este archivo tiene cuatro conjuntos de datos, los tres primeros son conocidos como dimensiones compartidas y son usados para definir la forma de la variable, el primero nombrado “lon”, almacena los 720 valores de longitudes para cada índice de esta variable coordenada, el segundo nombrado “lat”, contiene los 360 valores de las latitudes para cada índice de esta variable coordenada. Mientras que “time” contiene todos los valores de tiempo en que han sido tomados los datos, los cuales son expresados en días a partir de enero de 1900.

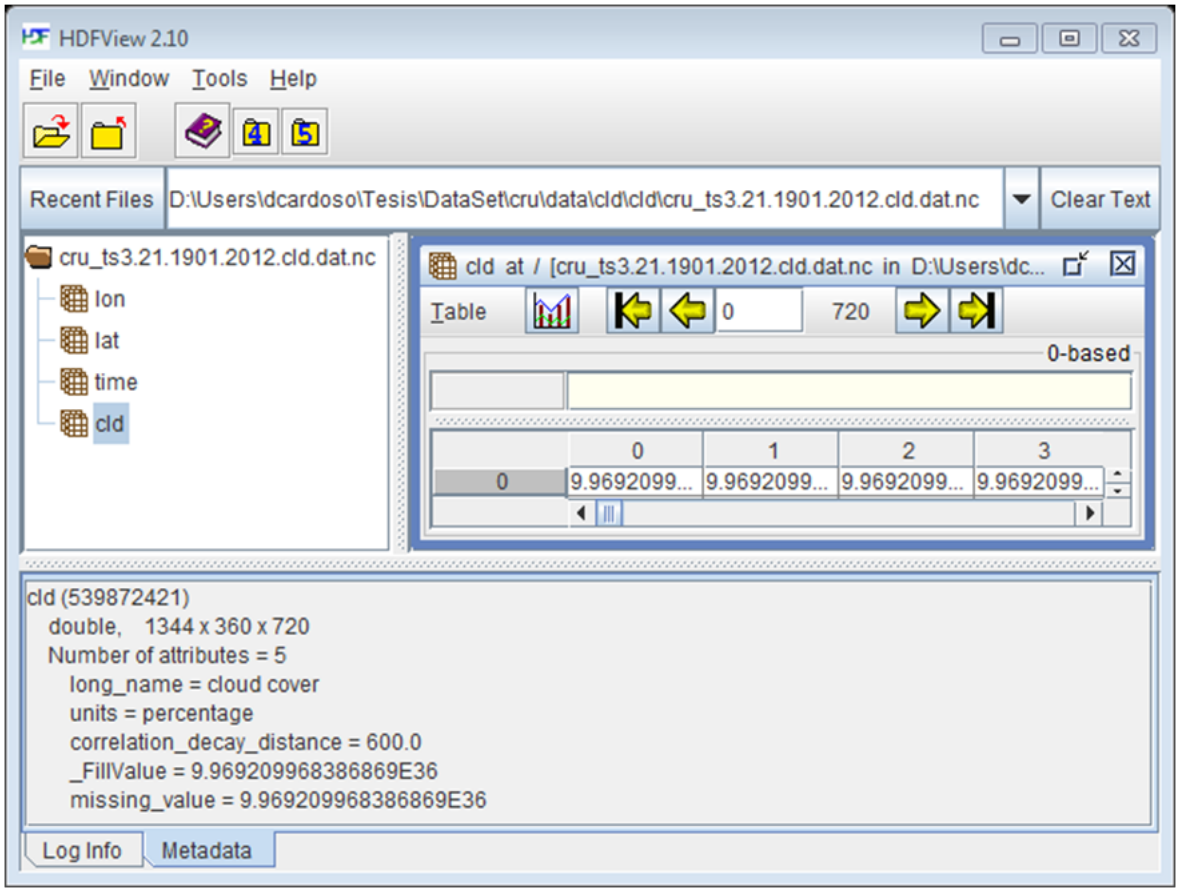

Figura 6.7 Estructura de la variable climática cld en formato NetCDF.

El conjunto de datos "cld", contiene los datos para la variable cobertura nubosa. Esta variable es almacenada como un contenedor de datos de tres dimensiones $L * M * N$ (donde L representa los 1344 momentos de tiempo, M los 360 valores de latitudes y N los 720 valores de longitudes respectivamente).

La idea de este caso de estudio es ejemplificar la utilización del módulo realizado para automatizar el proceso de conversión de dichos datos a un archivo HDF, con la estructura requerida por el módulo de visualización científica, desarrollado en el capítulo anterior.

Para llevar a cabo esta tarea se utilizó el Modelador Gráfico de Sextante, como se observa en la figura 6.8, donde se realizó todo el proceso de configuración para la transformación. Se aplicó 10 veces el algoritmo NetCDFcrutoHDF para cada una de las variables presentadas anteriormente en la tabla 5.2, que componen el archivo CRU TS3.21. Cada uno de estos algoritmos tiene como salida el mismo archivo HDF, la idea es almacenar todas las variables en un solo HDF. Además, estos algoritmos crean un conjunto de datos por cada una de las variables, que almacena sus valores máximos y mínimos, y otro que se crea una sola vez por el primer algoritmo: el conjunto de datos "Index" que es el mismo para todas las variables. Para este caso de estudio no se suministró un nivel de compresión; esto implica que el archivo creado ocupa mayor espacio en disco, pero a la vez agiliza considerablemente la escritura y lectura del archivo HDF.

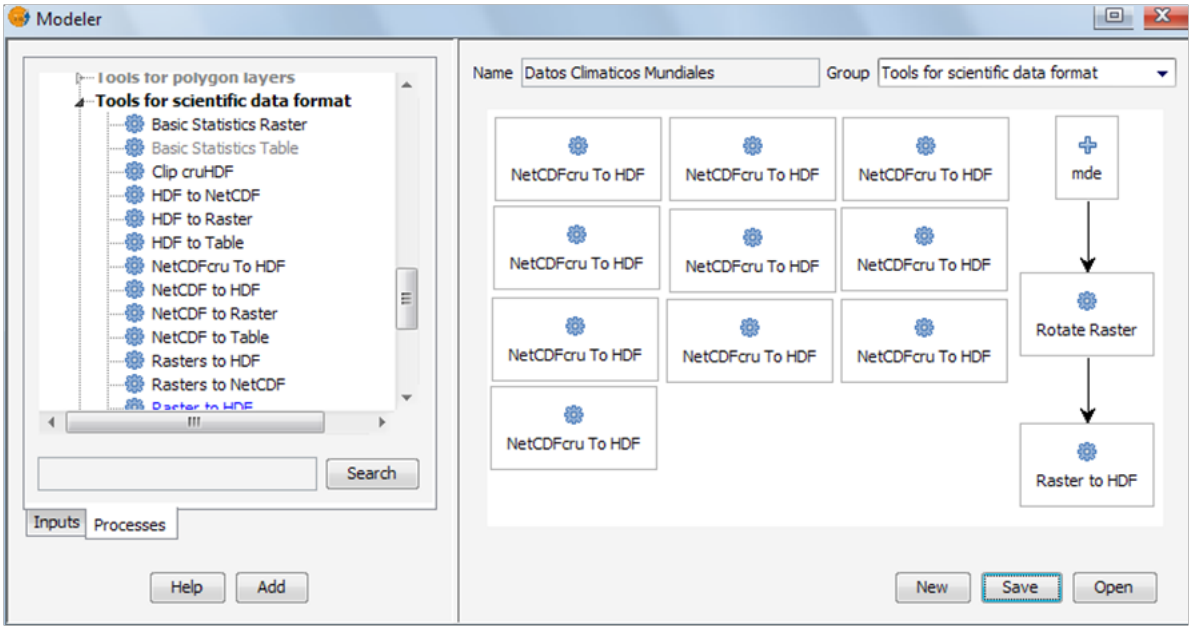

Figura 6.8 Flujo de trabajo para convertir el *dataset* CRU TS3.21 a HDF.

El algoritmo RastertoHDF fue utilizado para crear el modelo digital de elevaciones dentro del archivo HDF, para esto se tuvo que cargar previamente un raster con el modelo digital de elevaciones del planeta de 720 por 360 en una vista de gvSIG. Este raster se tuvo que rotar con el algoritmo RotateRaster, que rota 90 grados en contra de las manecillas del reloj; de esta forma el conjunto de datos "mde" queda como un *raster* de 360 por 720. En este proceso de transformación se respetó el tipo de dato *double* con el cual fueron creadas las variables originalmente por el CRU en formato NetCDF.

La ejecución de todo este modelo puede demorar un tiempo considerable, pero los algoritmos están diseñados para realizar esta tarea de una manera escalada.

Hasta este punto se ha generado un archivo HDF5, que contiene diez variables climáticas para toda la superficie terrestre. El archivo llamado "Planeta tierra.h5", tiene un poco más de 7GB de datos y posee la estructura necesaria para ser utilizado por el módulo de visualización científica de gvSIG.

El siguiente paso para finalizar este caso de estudio, consiste en recortar el archivo "Planeta tierra.h5" mediante la aplicación del algoritmo ClipcruHDF, con el objetivo de generar un nuevo HDF con los datos de una región de interés. En este caso se comentó que se deseaba generar un conjunto de datos para realizar análisis exploratorio sobre datos climáticos de la península Ibérica. El algoritmo ClipcruHDF se ha utilizado de conjunto con el módulo de visualización científica, el cual provee una herramienta que le muestra al usuario las coordenadas del punto superior izquierdo y el punto inferior derecho de una región seleccionada sobre el mapa de fondo. Estas coordenadas son pasadas como parámetros del algoritmo ClipcruHDF. Se debe especificar el camino del HDF de entrada y el camino del HDF de salida.

Este algoritmo se puede ejecutar desde la caja de herramientas de Sextante, de una manera

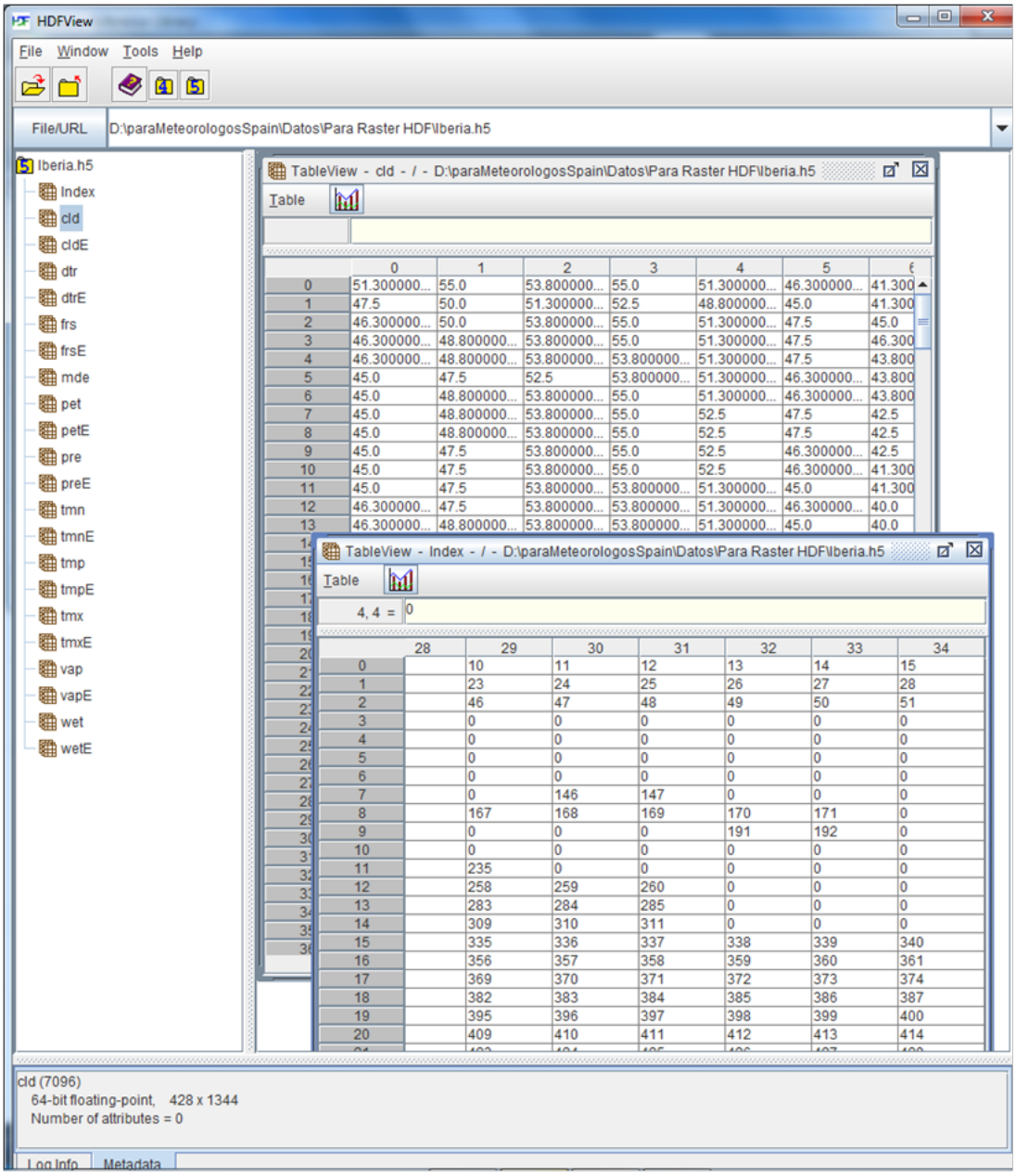

Figura 6.9 HDF con datos climáticos de la península Ibérica.

diferente a los ejecutados desde el Modelizador Gráfico mostrados en el ejemplo anterior. El HDF de salida se puede nombrar “Iberia.h5”. Después de la ejecución del algoritmo, el archivo contiene una estructura como la que se muestra en la figura 6.9.

Básicamente, este algoritmo hace una selección de los datos de cada variable que se encuentra dentro de la región seleccionada. Se calculan y se crean nuevamente los conjuntos de datos que almacenan los extremos con los máximos y mínimos por variable de los datos seleccionados. Los conjuntos de datos de las variables almacenan la cantidad de filas con información que se encontraron en el recuadro de recorte. Se calcula nuevamente el conjunto de dato “Index” para los datos seleccionados y se recorta el modelo digital de elevaciones. En la figura 6.10, se muestra el modelo digital de elevaciones de la península Ibérica luego de cargar con el módulo

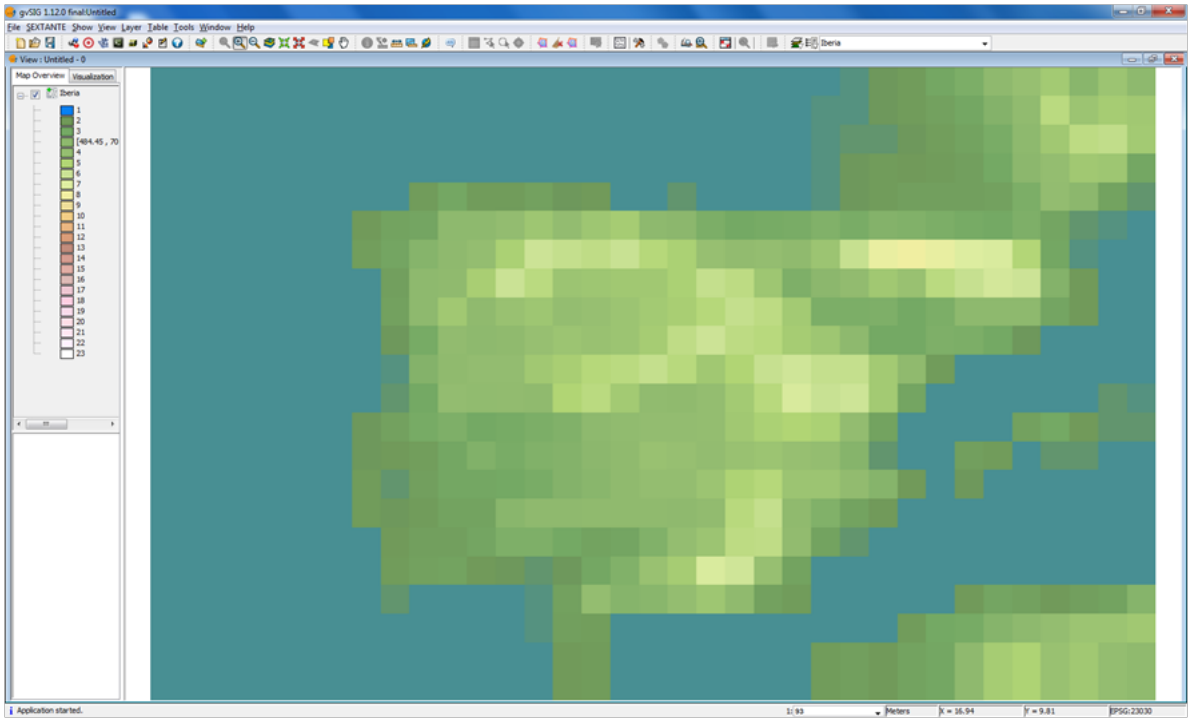

Figura 6.10 Modelo digital de elevaciones de la península Ibérica.

de visualización científica el HDF generado.

El tiempo de ejecución de este algoritmo para la región seleccionada no es alto, pero de manera general este tiempo depende de cuán grande sea la región de estudio seleccionada y qué nivel de compresión presentan los datos en el archivo HDF de entrada, del cual se quiere tomar el subconjunto de datos.

6.4. Conclusiones parciales

En este capítulo se describió un conjunto de herramientas y algoritmos que permiten la manipulación de los formatos de datos científicos HDF y netCDF en sistemas de información geográfica. En particular las herramientas han sido implementadas como una extensión de la biblioteca Sextante. Estas herramientas pueden ser utilizadas en cualquier sistema de información geográfica basado en Java que permita la integración de esta biblioteca.

La extensión desarrollada en este capítulo es de gran utilidad para la transformación de formatos de datos comunes en los sistemas de información geográfica para formatos de datos científicos y viceversa. Además algunos de los algoritmos están diseñados para facilitar la creación automática de conjuntos de datos con la estructura necesaria que requiere el módulo de visualización científica de gvSIG descrito en el capítulo anterior. Se ha desarrollado un caso de estudio que evidencia la viabilidad en la utilización de estas herramientas para crear grandes conjuntos de datos que pueden ser analizados mediante visualizaciones en sistemas de información geográfica.

7 Caso de estudio integrador y validación de los resultados

Este capítulo se realiza la validación de los resultados de esta tesis mediante la presentación de un caso de estudio integrador donde se han utilizado las herramientas desarrolladas para estudiar los tipos de clima en la península ibérica. En la sección 7.2 se ha aplicado el método de expertos para validar los aportes de este trabajo.

7.1. Caso de estudio integrador climatología de la península ibérica

En esta sección se pone de manifiesto y ejemplifica la gran utilidad que tiene para la climatología un grupo de nuevas opciones gráficas implementadas en gvSIG sobre formatos científicos de datos climáticos masivos. A partir de diferentes casos de estudios climáticos se demuestra la aplicabilidad de las herramientas diseñadas utilizando como ejemplo la distribución espacial de los diferentes tipos de clima de la península ibérica. No se trata de realizar un estudio exhaustivo, sino de tomar algunos climas o subtipos de clima que puedan, de forma clara, poner de manifiesto la utilidad y versatilidad de estos gráficos que se han desarrollado en la realización de estudios climáticos.

Todos los instrumentos que apoyen el desarrollo de estudios del clima en un territorio con el objetivo de mejorar el conocimiento sobre los mismos serán una excelente herramienta, valiosa e indispensable, como vehículo de transmisión de información del clima para dar apoyo al desarrollo socioeconómico. Por medio de la caracterización actualizada del clima se puede dar respuesta a algunos de los requerimientos más apremiantes que se plantea la gestión medioambiental.

La mejora existente hoy en la disponibilidad de datos y en el conocimiento de los cambios en los extremos climáticos es fundamental para apoyar la detección de los cambios y variaciones que está experimentando el clima debido a las influencias antropogénicas sobre la atmósfera. Las representaciones gráficas son un importante instrumento de referencia para el seguimiento de la variabilidad y del cambio climático.

A partir de diferentes ejemplos se ponen de manifiesto algunas de las potencialidades de estas herramientas gráficas de gran utilidad, que básicamente pueden resumirse en dos:

1. El manejo de grandes bases de datos climáticas.
2. Análisis espacio-temporal de variables climáticas.

7.1.1. El manejo de las grandes bases de datos climatológicas

La disponibilidad masiva de datos científicos existentes en la web supone un hito en el ámbito de la difusión y la reutilización de los mismos, lo que genera un crecimiento exponencial del impacto actual de los avances científicos en general (Gray *et al.*, 2005). En el ámbito de la climatología estas nuevas bases de datos climáticas en formatos científicos (NetCDF y HDF) suponen una revolución en el seno de la disciplina, particularmente si tenemos en cuenta la dificultad histórica en la disponibilidad de datos climáticos, incluso en ámbitos territoriales desarrollados (Harris *et al.*, 2014). Por un lado, el limitado y difícil acceso a la información climática ya mencionado ha sido una de las características que ha frenado en gran medida el desarrollo de la Climatología hasta hace pocos años. Aún hoy en España los datos correspondientes a las series de registros de las variables meteorológicas de las diferentes redes de la Agencia Estatal de Meteorología española (Aemet), no pueden obtenerse salvo pago de tasas. No existe ningún tipo de información libre a la que los usuarios puedan tener acceso, lo que contrasta con la nuevas directrices y tendencias globales como la directiva Inspire en Europa a las iniciativas del Open Consortium (OCG). Un segundo problema ha sido un freno hasta hace poco, la mala, escasa o nula información climatológica en algunas zonas del planeta por la inexistencia de observatorios meteorológicos o la mala calidad de los registros; como ejemplo, el caso de extensas zonas del continente africano, Sudamérica o las zonas montañosas. La escasa cobertura espacial ha hecho que hasta la aparición de los satélites meteorológicos el clima de algunas áreas del planeta fuese prácticamente desconocido. A esto hay que añadir también que la extensión temporal de las series varía enormemente, por lo cual resulta difícil remitir a un período común los estudios.

Otro problema importante asociado a las bases de datos climáticas son los problemas de calidad y homogeneidad. Una serie climática se dice que es homogénea y de calidad en el tiempo, cuando refleja exclusivamente las variaciones experimentadas por la atmósfera y no cambios asociados a la medición de la variable. Las causas más habituales de errores y de falta de homogeneidad son cambios instrumentales tales como relocalizaciones de estaciones de medición, cambios en las prácticas observacionales, cambios en los alrededores de las estaciones en cuanto a los usos o el tipo de cubierta de suelo, efecto urbano, etc. Por ello los datos climáticos deben ser sometidos a estrictos controles y procesos, con la finalidad de etiquetar datos potencialmente erróneos, validarlos tras consulta o rechazarlos antes de usarlos y, por último, someterlos a pruebas estadísticas de homogeneidad. Todo ello es un proceso largo y complejo que explica la limitada disponibilidad de registros de calidad.

Tabla 7.1 Limitaciones de la información tradicional y ventajas de las nuevas fuentes de información.

TRADICIONALMENTE	PRESENTE
Cobertura irregular	Cobertura muy regular, proliferación de *grids*
Muy diversos alcances temporales	Largo alcance y pocas lagunas
Escaso control de calidad y homogeneidad	Calidad más testada
No globales	Cobertura global
Inaccesibles o caras	Libre acceso
No todas las variables disponibles	Diversidad de variables

Pese a que la atmósfera ha sido concienzuda y regularmente monitorizada desde mediados del siglo XIX para la mayor parte del globo y con anterioridad en determinadas regiones, actualmente se dispone de menos registros y de peor calidad de los que se requieren para estudiar el clima y, más importante aún hoy, detectar y atribuir eventos al cambio climático. La accesibilidad a registros de calidad sigue limitada pese a las llamadas de agencias internacionales (e.g. UNFCCC, GEOSS, GCOS, etc.) o la resolución 40 de la Organización Meteorológica Mundial por problemas de escasez de recursos humanos y financieros, de propiedad de datos (patrimonio nacional) y de interoperabilidad de redes.

Así puede entenderse mejor por qué hoy en día la disponibilidad de numerosas y diferentes bases de datos completas, tanto espacial como temporalmente en formatos científicos supone una revolución para los estudios climáticos y atmosféricos. Podemos resumir en el siguiente cuadro las limitaciones de la información tradicional en climatología y las ventajas de estas nuevas fuentes de información (obsérvese la tabla 7.1):

A pesar de que todos estos problemas en los datos relacionados, básicamente, con la calidad, disponibilidad y accesibilidad los solucionan en gran medida las *nuevas* grandes bases de datos disponibles, lo que supera las limitaciones de información tradicionales en climatología, hay que decir que estas también presentan algunas limitaciones y dificultades. En primer lugar, los formatos suelen ser aún de muy difícil manejo y gestión, lo que limita su utilización. En segundo, en algunos casos las referencias espaciales o temporales no son coincidentes entre unas y otras.

Por esta razón, la posibilidad de utilización de estas bases de datos y sus formatos de una forma sencilla y gráfica en entornos de gestión de la información geográfica supone una gran contribución que facilita enormemente el uso práctico y las posibilidades de análisis exploratorio de estas completas y valiosas fuentes de datos.

7.1.2. El análisis espacio-temporal de variables climáticas

Como se ha comentado, disponer de información climatológica continua en el espacio que permita realizar estudios completos de territorios, regiones o países es difícil, tanto por los motivos antes comentados, como por las técnicas de análisis que requieren. Un paso previo imprescindible es el análisis exploratorio de la información y la caracterización espacial, algo que

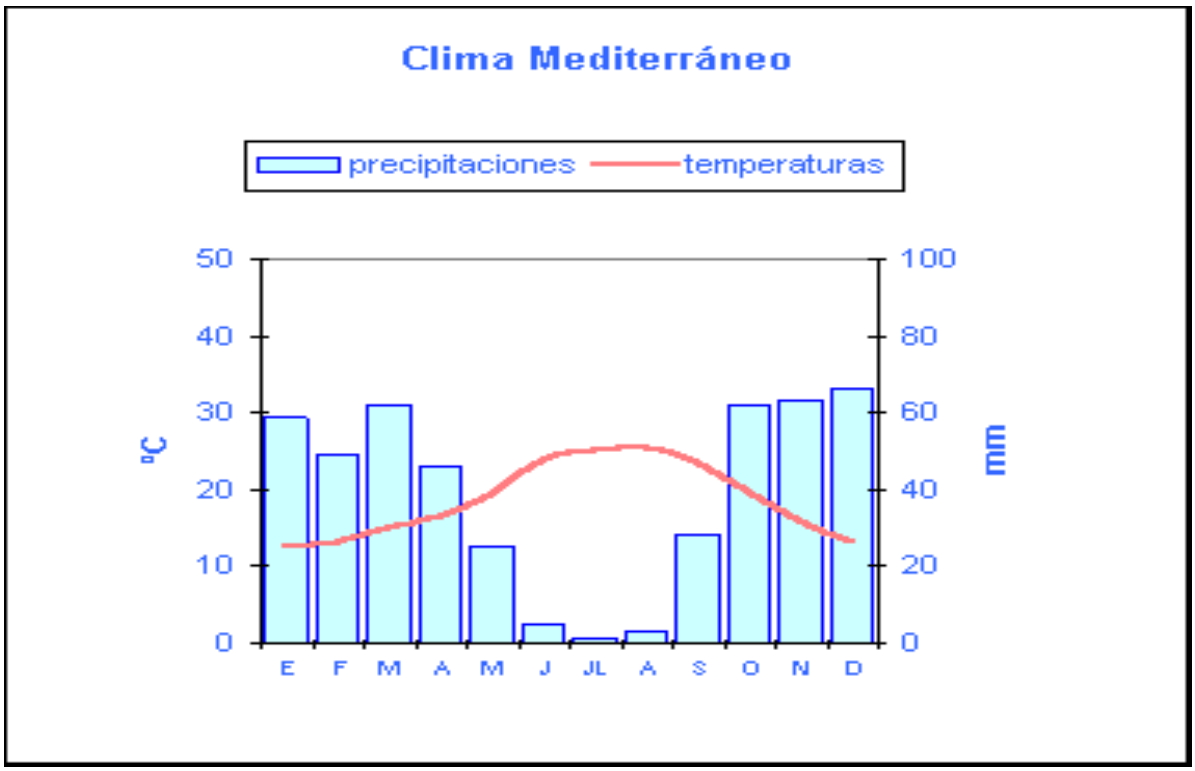

Figura 7.1 Climograma del Mediterráneo.

las nuevas herramientas implementadas permiten realizar fácilmente, como se puede observar en los casos de estudio realizados en esta sección. La posibilidad de representar tanto múltiples variables o de seleccionar las que interesen en diferentes tipos de gráficos, permiten poner de manifiesto, las diferencias espaciales y de comportamiento entre unos lugares y otros, cubriendo además, todo el territorio seleccionado.

Otro de los aspectos más interesantes y necesarios en el estudio del clima es su dimensión temporal. En el clima los fenómenos interesan por su duración o persistencia, por su repetición y variaciones. Por esta razón se caracterizan sus valores medios y las probabilidades de producirse valores extremos en los diversos elementos climáticos.

El análisis gráfico de series temporales completas, permite superar una limitación tradicional en la representación de las variables climáticas basada en valores medios, muchas veces poco representativos, que únicamente son capaces de indicar el comportamiento del régimen anual. El tipo de representación más frecuente, clásico y estándar en climatología, ha sido el climogramas de Walter y Gaussen que únicamente utiliza la precipitación y temperatura media a lo largo del año (obsérvese la figura 7.1).

Los climogramas sirven para una descripción muy general del clima sin permitir el estudio de otros aspectos importantes del mismo como su variabilidad. Esto es especialmente importante en el caso de climas muy variables, como el mediterráneo, caracterizados por una gran irregularidad espacial y temporal, y donde la presencia de valores extremos tiene gran importancia por los riesgos asociados (inundaciones y sequías). Por esta razón la visualización de extremos es otro

de los aspectos fundamentales en el estudio del clima que se aborta en casos de estudio.

La mejora existente hoy en la disponibilidad de datos y en el conocimiento de los cambios en los extremos climáticos, es fundamental para apoyar la detección de los cambios y variaciones que experimenta el clima, debido a las influencias antropogénicas sobre la atmósfera. Resulta difícil el examen temporal de las series históricas de datos climáticos en donde se pueda apreciar y caracterizar su evolución, valores extremos, etc. En este sentido, la representación gráfica de estos datos es un paso básico y esencial en la descripción de la dimensión temporal del clima, lo que, unido a las ventajas de continuidad espacial y utilización de variables combinadas o individualmente, según se elija, vuelve a ser un gran avance como, de igual manera, se muestra en los casos prácticos. La representación espacio-temporal de las variables climáticas para un territorio, con la utilización de toda la información disponible, es algo completamente novedoso, que puede cambiar la descripción e interpretación en detalle de muchas descripciones climatológicas.

7.1.3. La climatología de la península ibérica

Todos los aspectos comentados anteriormente como el análisis combinado de variables, el poder incluir variables no contempladas en las clasificaciones climáticas habituales, la posibilidad de estudio de los matices espaciales que se ejemplifican, son analizados a partir de las opciones gráficas desarrolladas. Con toda seguridad, permiten el estudio con mayor detalle de todos estos aspectos del clima, y contribuyen a redefinir tipos y subtipos climáticos.

Para los casos de estudio se realiza una selección de tipos de clima de la península ibérica, que son adecuados para demostrar la aplicabilidad de este trabajo. Para ello, se utilizan las tipologías descritas y presentadas en el atlas climático de la península ibérica, publicado en 2011 por la Agencia Estatal de Meteorología de España (Aemet) (AEMET, 2011). Se trata de un documento de referencia, tanto en el panorama nacional como en el europeo, que sirve de soporte a las actividades de varias entidades públicas y privadas, así como a los ciudadanos de Portugal y España. El Atlas climático de España utiliza la clasificación climática de Köppen que, a pesar de haberse realizado hace unos 100 años, continúa siendo una de las más utilizadas para definir los climas del mundo.

Clasificación Climática de Köppen

Vladimir Köppen propone una clasificación climática en la que se tiene en cuenta tanto las variaciones de temperatura y humedad como las medias de los meses más cálidos o fríos, y lo más importante, destaca las consecuencias bioclimáticas. Köppen publica su clasificación definitiva en 1936. En 1953, dos de sus alumnos, Geiger y Pohl revisan la clasificación, por lo que también se conoce como clasificación de Köppen-Geiger-Pohl (Strahler y Strahler, 1992). En la clasificación, el clima se divide en grupos climáticos, subgrupos y subdivisiones. Los

Tabla 7.2 Grandes tipologías climáticas de la clasificación climática de Köppen.

A	Climas lluviosos tropicales	La temperatura media es superior a los 18 °C. Carecen de invierno. El mes más frío tiene una temperatura superior a los 18 °C.
B	Climas secos	La evaporación excede las precipitaciones. Siempre hay déficit hídrico.
C	Climas templados y húmedos. Mesotérmicos	Temperatura media del mes más frío es menor de 18 °C y superior a -3 °C y al menos un mes la temperatura media es superior a 10 °C. Poseen verano e invierno.
D	Climas boreales o de nieve y bosque. Microtérmicos	La temperatura media del mes más frío es inferior a -3 °C y la del mes más cálido superior a 10 °C
E	Climas polares o de nieve	La temperatura media del mes más cálido es inferior a 10 °C y superior a 0 °C. No tienen un verdadero verano.
F	Clima de hielos perpetuos	La temperatura media del mes más cálido es inferior a 0 °C

Tabla 7.3 Subtipos climáticos.

S	Semiárido (Clima de Estepa)	Sólo para climas de tipo B. De 380 a 760 mm.
W	Árido (Clima Desértico)	Sólo para climas de tipo B. Menos de 250 mm.
f	Húmedo (sin estación seca)	Solo para climas de tipo A, C y D. Precipitación todo el año.
m	Húmedo (Clima Bosque lluvioso) Pluviisilva	Solo para climas de tipo A, con una corta estación seca. Tipo monzónico.
W	Estación seca en invierno	Sol en posición baja
s	Estación seca en verano	Sol en posición alta

grupos climáticos se establecen en función de la temperatura mensual media, y estas grandes tipologías aparecen representadas cada una por una letra mayúscula. De estas tipologías en la península Ibérica se encuentran representadas dos de ellas, la C, correspondiente a climas templados y húmedos (Mesotérmicos) y la B correspondiente a los climas secos (obsérvese tabla 7.2).

A partir de estos grandes grupos climáticos se matizan los subtipos utilizando también letras para definir subgrupos, básicamente diferenciados por la humedad. Los dos primeros se escriben con mayúscula y el resto con minúscula (obsérvese la tabla 7.3).

De la combinación de grupos y subgrupos se obtienen doce tipos de clima básicos (obsérvese la tabla 7.4):

Köppen añadió una tercera letra a los distintos tipos ya mencionados. Las subdivisiones

Tabla 7.4 Tipos de clima básicos.

Af	Clima de selva tropical lluviosa	El mes más seco caen más de 600 mm de lluvia
Am	Clima monzónico	El mes más seco caen menos de 600 mm de lluvia
Aw	Clima de sabana tropical	Por lo menos hay un mes en el que caen menos de 600 mm
BS	Clima de estepa	Clima árido continental
BW	Clima desértico	Clima árido con precipitaciones inferiores a 400 mm
Cf	Clima templado húmedo sin estación seca	Las precipitaciones. del mes más seco son superiores a 300 mm
Cw	Clima templado húmedo con estación invernal seca	El mes más húmedo del verano es 10 veces superior al mes más seco del invierno
Cs	Clima templado húmedo con veranos secos	Las precipitaciones del mes más seco del verano es inferior a 300 mm y del mes más lluvioso del invierno 3 veces superior
Df	Clima boreal de nieves y bosque con inviernos húmedos	No hay estación seca
Dw	Climas boreales o de nieve y bosque con inviernos secos	Con una estación seca en invierno
ET	Clima de tundra	Temperatura media del mes más cálido es inferior a 10 °C y superior a 0 °C
EF	Clima de los hielos polares	La temperatura media del mes más cálido es inferior a 0 °C

dependen de características adicionales. Se expresan en minúscula en la tabla 7.5.

Estos climas tienen variantes en función de las subdivisiones, por lo que cada clima se expresa con tres letras. En esta clasificación, en realidad, no se puede hablar de regiones climáticas, aunque se hace de manera general, sino de qué tipo de clima hay en un lugar atendiendo a estos criterios.

Los climas de la península ibérica

El clima de España es muy variado debido a su posición latitudinal, rodeada de masas de aguas y por el relieve tan diverso del territorio. La variada orografía de España, así como su situación geográfica, en latitudes medias de la zona templada del hemisferio norte hace que el país tenga una notable diversidad climática. Los grandes tipos climáticos según Köppen definidos a escala mundial para la península ibérica corresponden a los grupos Cbf, correspondiente a lo que de forma general se conoce como climas oceánicos o atlánticos, y que se extienden por toda la zona norte limitada a grandes rasgos por la cornisa cantábrica. La mayor parte del territorio se sitúa en el tipo Csb, asociado de forma general a los climas mediterráneos y, finalmente,

Tabla 7.5 Subdivisiones de los tipos de clima propuestos por Köppen.

a	Con verano caluroso	El mes más cálido por encima de 22 °C (climas C y D)
b	Con verano cálido	El mes más cálido por debajo de 22 °C (climas C y D)
c	Con verano corto y fresco	Menos de 4 meses por encima de 10 °C (climas C y D)
d	Con invierno muy frío	El mes más frío por debajo de – 38 °C (solo climas D)
h	Caluroso y seco	Temperatura anual media superior a 18 °C (solo climas B)
k	Frío y seco	Temperatura media anual por debajo de 18 °C (solo climas B)

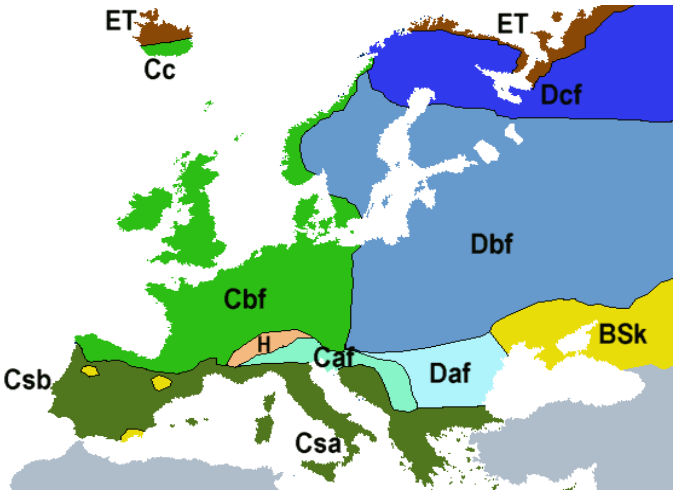

Figura 7.2 Tipos de clima en Europa.

aparecen zonas con el tipo Bsk, asociado a los climas de estepa, desérticos secos (obsérvese la Figura 7.2).

En conjunto la clasificación a escala global resulta demasiado general y claramente imprecisa para muchas zonas de España, por poner un ejemplo, la última tipología Bsk, que asocia en un mismo subgrupo los climas de montaña y los climas desérticos del sudeste peninsular. Por esta razón la Aemet ha mejorado la precisión espacial y la definición de subtipos con la utilización de la clasificación de Köppen para la península ibérica lo cual ha enriquecido notablemente la caracterización climática de ésta (obsérvese la figura 7.3)

En el atlas de la península ibérica se pasa de tres tipologías a 12 subgrupos climáticos, lo cual como se ha mencionado matiza enormemente y refleja con mayor exactitud su enorme diversidad climática.

Tradicionalmente, se han clasificado cuatro grandes climas en la península ibérica: oceánico; mediterráneo (con algunas variaciones); desérticos, y de montaña, y cada uno cada uno influye en un área geográfica claramente delimitada. Estos cuatro tipos corresponderían a grandes rasgos

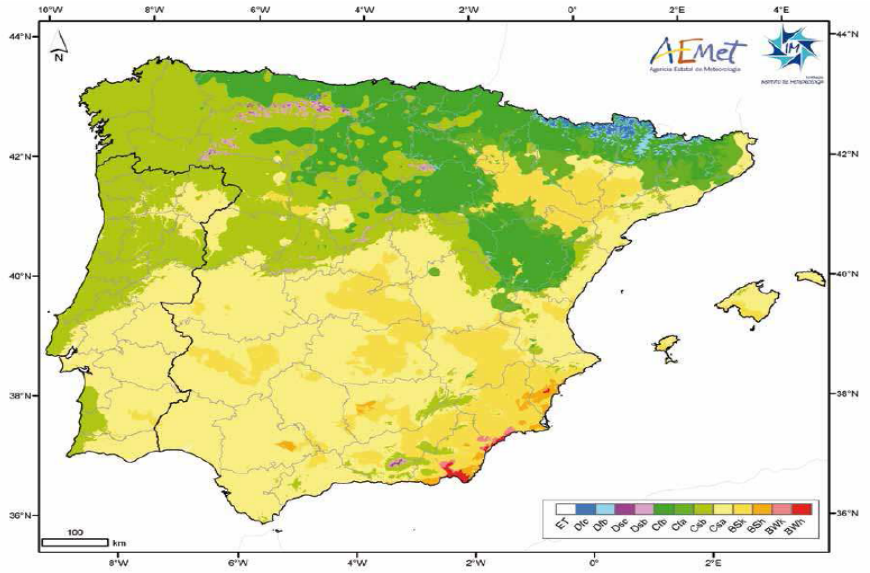

Figura 7.3 Clasificación climática de Köppen-Geiger en la península ibérica e islas Baleares

Tabla 7.6 Tipos de climas y rasgos de la península ibérica.

Tipo de clima general	Clasificación de Köppen
Oceánico	Cfa /Cfb
Mediterráneo	Csa/Csb
Montaña	Dfc/Dfb
Desérticos	BSk/BWk/BWh
Subtropical (Canarias)	No representado

a la clasificación que se muestra en la tabla 7.6 y que se detallan en las tablas 7.7, 7.8, 7.9 y 7.10.

7.1.4. Caso de estudio.

Los casos de estudio que se han diseñado para la evaluación de las herramientas gráficas implementadas están relacionados con la caracterización de las tipologías climáticas de la península ibérica. A partir de los criterios de clasificación y estudios climatológicos mencionados en los apartados anteriores se han formulado un grupo de casos que han sido utilizados para comprobar la utilidad de los métodos y herramientas desarrollados en esta tesis.

La caracterización de las tipologías climáticas de la península ibérica.

El objetivo en este caso de estudio es valorar y demostrar la utilidad de las herramientas gráficas desarrolladas, aplicadas a la multiserie climática de la península ibérica, en la definición, caracterización e interpretación de las tipologías climáticas presentes en la misma.

Tabla 7.7 Los climas mediterráneos

Los climas mediterráneos	El clima mediterráneo es el que predomina en España, ya que se extiende a lo largo de todo el litoral mediterráneo, el interior de la península y el archipiélago balear. Sin embargo, existen considerables diferencias entre unas zonas y otras, lo que da lugar a varias subdivisiones.
- Csa (templado con verano seco y caluroso)	Es la variedad de clima que abarca una mayor extensión de la península ibérica y baleares; ocupa aproximadamente el 40 por ciento de su superficie. Se extiende por la mayor parte de la mitad sur y de las regiones costeras mediterráneas, a excepción de las zonas áridas del sureste. Correspondería a lo que tradicionalmente se ha denominado clima mediterráneo con invierno frío o "mediterráneo continentalizado". Se localiza en la Meseta Ibérica, la depresión del Ebro, parte del Guadalquivir y la zona del norte de la provincia de Alicante. Se caracteriza por tener unas temperaturas muy extremas, entre 25 °C y los -13 °C. Los inviernos son largos y muy fríos, donde las temperaturas mínimas pueden bajar hasta los -5 grados o más, y los veranos muy calurosos, donde todos los años se sobrepasan los 35 grados, e incluso los 40 grados en algunas ocasiones. Además, las precipitaciones son escasas, en torno a los 400 mm, y aparecen en forma de tormenta en los meses de julio y agosto.
- Csb (templado con verano seco y templado)	Abarca la mayor parte del noroeste de la península, así como casi todo el litoral oeste de Portugal continental y numerosas áreas montañosas del interior de la Península.
- Cfa (templado sin estación seca y verano caluroso)	Se observa principalmente en el noreste de la Península, en una franja de altitud media que rodea los Pirineos y el Sistema Ibérico.
- Cfb (templado sin estación seca y verano templado)	Se localiza en la región cantábrica, en el sistema ibérico, parte de la meseta norte y gran parte de los Pirineos, excepto las áreas de mayor altitud.

Tabla 7.8 Los climas desérticos

- BWh (Desierto cálido) y BWk (Desierto frío)	Se localizan en pequeñas áreas del sureste de la península ibérica, en las provincias españolas de Almería, Murcia y Alicante, lo que coincide con los mínimos pluviométricos peninsulares. Las lluvias son extremadamente escasas, menos de 300 mm al año, lo que convierte estas zonas en áreas muy áridas, y son frecuentes los períodos largos de sequía. Las temperaturas son semejantes a las del mediterráneo típico, aunque el calor en verano suele ser más intenso. Las zonas frías se localizan en las mayores altitudes de estos ámbitos.
- BWh (Desierto cálido) y BWk (Desierto frío)	Se localizan en pequeñas áreas del sureste de la península ibérica, en las provincias españolas de Almería, Murcia y Alicante, lo que coincide con los mínimos pluviométricos peninsulares. Las lluvias son extremadamente escasas, menos de 300 mm al año, lo que convierte estas zonas en áreas muy áridas, y son frecuentes los períodos largos de sequía. Las temperaturas son semejantes a las del mediterráneo típico, aunque el calor en verano suele ser más intenso. Las zonas frías se localizan en las mayores altitudes de estos ámbitos.
- BSh y BSk (Estepa fría)	En España se extienden ampliamente por el sureste de la Península y valle del Ebro y, en menor extensión, en la meseta sur y Extremadura.

Tabla 7.9 El clima oceánico

El clima oceánico	Este clima se extiende por todo el norte y noroeste de la Península, desde los Pirineos hasta Galicia. Se caracteriza por la abundancia de lluvias, que suelen superar los 1000 mm, repartidas de manera regular a lo largo del año. Por esa razón, el paisaje es muy verde. Las temperaturas suelen ser suaves, debido a la cercanía del mar: en invierno, oscilan entre los 12 °C y los 15 °C y en verano, rondan los 20-25 °C. Las ciudades representativas son San Sebastián, Santander, Foz, Vigo y Oviedo. De todas formas, y por ser principalmente ciudades costeras, la humedad suele intensificar las temperaturas mínimas y máximas, sobre todo en el sur de Galicia.
- Cfa (Templado sin estación seca y verano caluroso)	Se observa principalmente en el noreste de la Península, en una franja de altitud media que rodea los Pirineos y el Sistema Ibérico.
- Cfb (Templado sin estación seca y verano templado)	Se localiza en la región cantábrica, en el Sistema Ibérico, parte de la meseta norte y gran parte de los Pirineos, excepto las áreas de mayor altitud.

Tabla 7.10 El clima de montaña

El clima de montaña	Corresponde a los climas Fríos (Tipo D) que aparecen en los grandes sistemas montañosos como los Pirineos, el Sistema Central, el Sistema Ibérico, la cordillera penibética y la cordillera cantábrica. Los inviernos son muy fríos, y los veranos frescos. Las precipitaciones son muy abundantes a medida que aumenta la altitud y, en general, en forma de nieve. Las vertientes de las montañas que miran al norte son más frías. La temperatura media del mes más frío en este tipo de clima es inferior a 0 °C y la temperatura media del mes más cálido es superior a 10 °C.
- Dsb (frío con verano seco y templado) y Dsc (frío con verano seco y fresco)	Se localizan en pequeñas áreas de alta montaña de la cordillera cantábrica, sistema ibérico, sistema central y sierra nevada.
- Dfb (frío sin estación seca y templado) y Dfc (frío sin estación seca y fresco)	Se observan en áreas de alta montaña de los Pirineos y en algunas pequeñas zonas de alta montaña de la cordillera cantábrica y del sistema ibérico.

Para ello se ha seleccionado una serie de puntos en la península ibérica, distribuidos espacialmente de tal manera que permitan abarcar los subtipos climáticos mediterráneos más representativos de la misma según la clasificación de Köppen (obsérvese la figura 7.4).

La segunda fase consiste en la realización de un proceso de interpretación visual de imágenes y gráficos que permiten la identificación de las características esenciales de cada uno de los subtipos climáticos seleccionados. De esta manera se ha puesto especial énfasis en aquellos elementos que aportan información y valor añadido a los valores sintéticos absolutos que caracterizan los subtipos climáticos según la clasificación de Köppen.

Las herramientas gráficas desarrolladas que presentan mayor potencial de interpretación en la caracterización del comportamiento climático de las variables presentes en la multiserie de trabajo han sido las siguientes:

- Segmentos de círculo: herramienta muy útil para aislar, discriminar y contrastar mediante color, trama y distribución, los comportamientos diferenciales entre variables y, paralelamente, para encontrar patrones gráficos que pueden ser asociados a subtipos climáticos característicos.
- Combo temporal: herramienta muy útil para captar de forma instantánea las características propias y diferencias en el comportamiento de las variables climáticas de cada subtipo en términos de dimensión diacrónica del mismo.
- Patrones recursivos: herramienta de contraste en la que la combinación de colores diferencial a la que pueden someterse las variables climáticas, permite caracterizar los

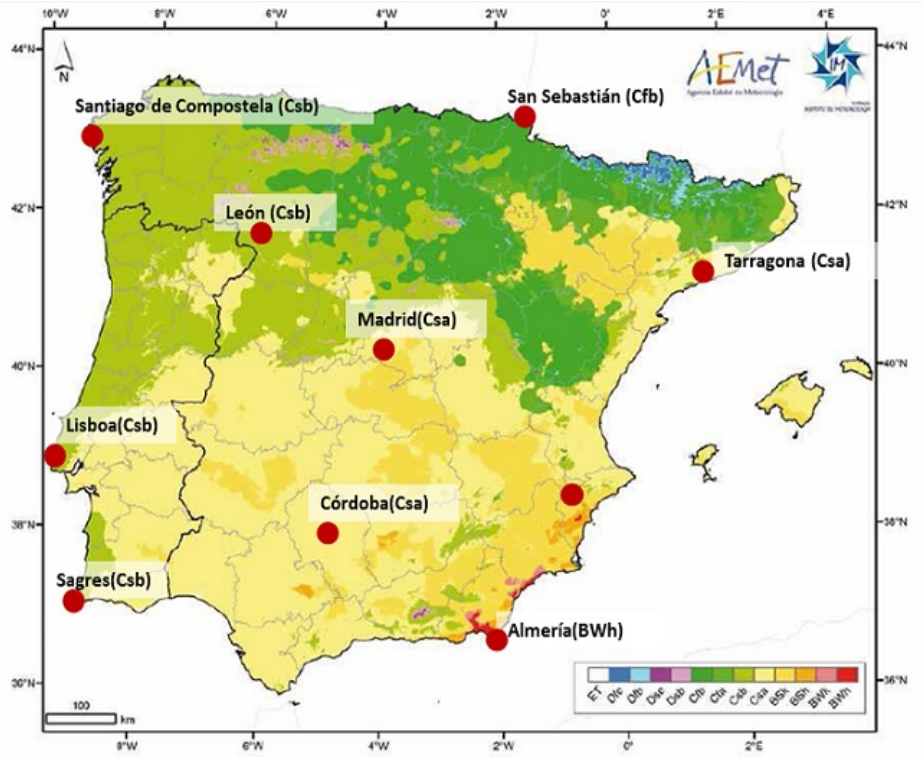

Figura 7.4 Selección de puntos con tipos de climas característicos de la península ibérica.

diversos patrones climáticos mendiante la definición de un patrón que organice los datos por diferentes estaciones, años, quinquenios, etc.

- Coordenadas de estrella: seguramente una de las herramientas con mayor capacidad para establecer patrones y que, además, es muy cercana a la climatología por compartir metodología constructiva con uno de los gráficos más utilizados en la disciplina como lo es la rosa de los vientos.
- Matriz de diagramas de dispersión: herramienta de visualización estadística clásica que permite observar las relaciones entre variables en los distintos subtipos climáticos de trabajo.

Los resultados obtenidos y las aportaciones de las técnicas de visualización seleccionadas para este primer caso de estudio han arrojado resultados de gran interés en los procesos de identificación y caracterización de subtipos climáticos. Algunos ejemplos de resultados se muestran a continuación.

a) Identificación de patrones climáticos a partir de la utilización de la herramienta visual segmentos de círculo (obsérvese la figura 7.5):

Tal y como se observa en la figura para las variables precipitación (pre), cobertura nubosa (cld) y frecuencia de días lluviosos (wet), todas ellas relacionadas con los procesos de génesis y acontecimiento de la precipitación y para las que se ha seleccionado una paleta de contraste en la que los valores cercanos al rojo indican valores bajos y los cercanos al azul indican valores más altos, puede observarse una óptima visualización de los patrones climáticos de esas variables

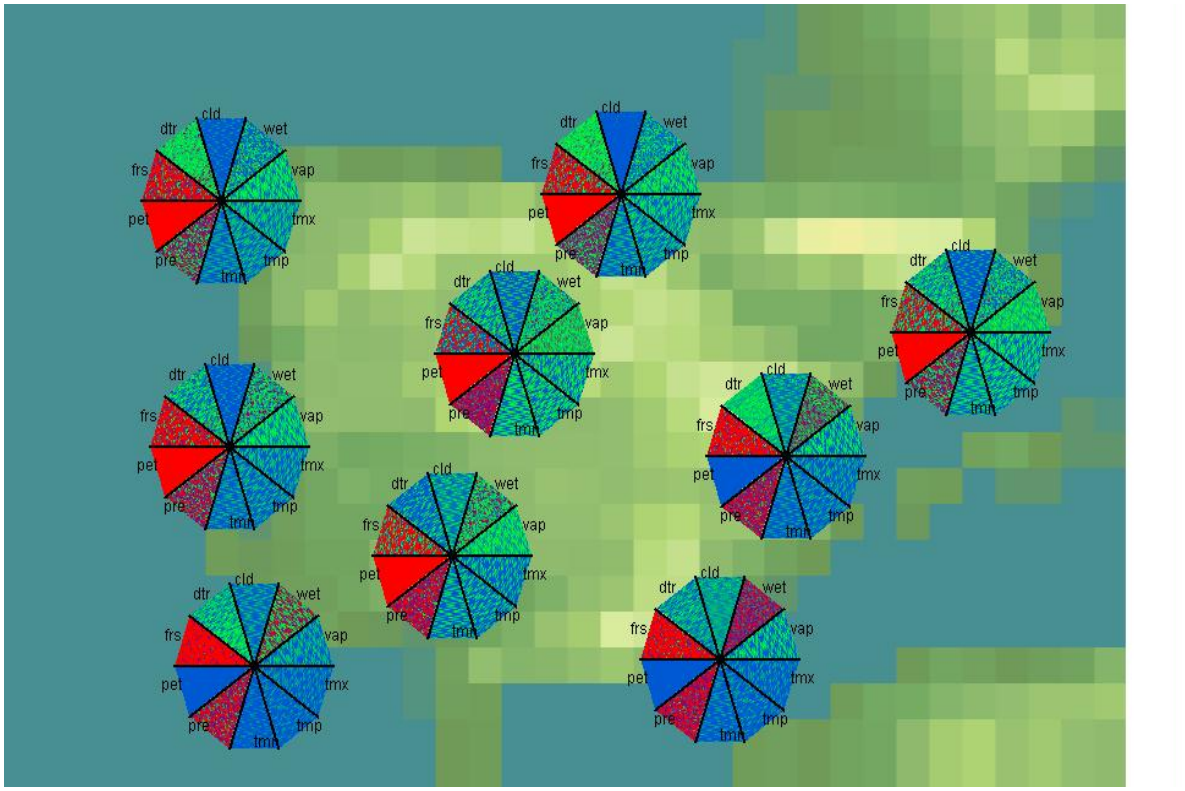

Figura 7.5 Segmentos de círculo sobre los puntos característicos de la península ibérica.

que pueden encontrarse en la península ibérica.

De esta forma se observa un gradiente muy claro NW-SE en términos de reducción tanto de los valores de precipitación, como especialmente de nubosidad y de humedad, reflejando perfectamente el tránsito de la Iberia húmeda a la Iberia seca y desértica. Es destacable que, al trabajar con el conjunto del vector mensual, la variable precipitación, aun en el punto coincidente con Santiago de Compostela, presenta alternancia de rojos y azules, respondiendo de esta manera la herramienta gráfica a la presencia de veranos menos húmedos como nota característica del clima mediterráneo oceánico de esta zona de la península ibérica. Aun así, el contraste visual con el punto más al SE (Almería) y más seco en términos de totales pluviométricos es muy significativo al ser este un dominio subdesértico (predominancia de los colores rojos tanto en precipitación como en frecuencia de días lluviosos).

Muy relevante es también la cobertura nubosa, con predominio de los azules intensos tanto en la zona SE como en la zona norte (San Sebastián), donde si bien los descensos en los niveles pluviométricos del verano son apreciables en la variable precipitación, no ocurre lo mismo en la variable nubosidad, en la cual los colores azules son los predominantes como consecuencia de meses veraniegos con menos lluvia pero muy nubosos como consecuencia de la afectación de estas zonas del frente polar que, aún desplazado a latitudes septentrionales, afecta también en estos meses a dichas zonas de la península ibérica, dejando menos precipitaciones, pero muchos días de nubosidad intensa.

b) Dimensionado del comportamiento temporal de las variables mediante técnica combo

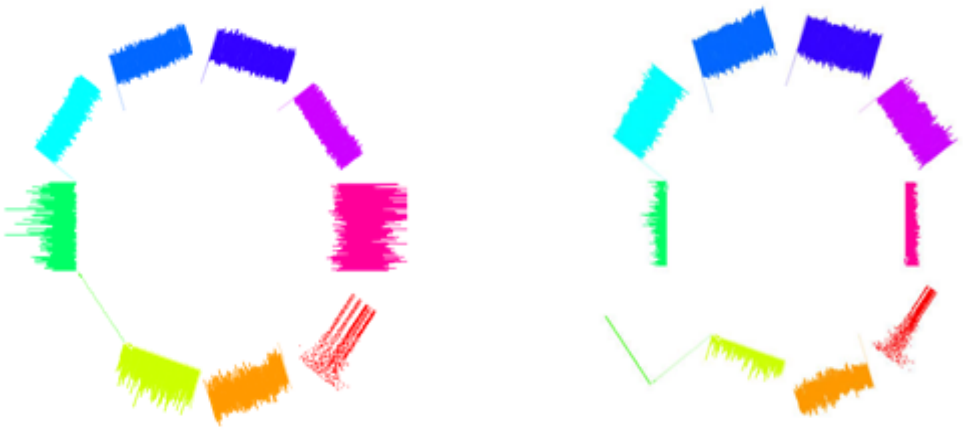

Figura 7.6 Combo temporal de Santiago de Compostela (izquierda) y de Almería (derecha). Las variables se organizan en el sentido de las manecillas del reloj siguiendo el orden cld, dtr, frs, pet, pre, tmp, tmn, tmx, vap y wet, comenzando por cld que corresponde con la secuencia de color rojo que se encuentra en la esquina inferior derecha de cada combo temporal.

temporal (obsérvese la figura 7.6).

Esta técnica permite una rápida distinción entre dos de los climas más contrastados de la península ibérica ya que permiten, de una forma automática, dimensionar el comportamiento de las variables básicas y encontrar patrones diferenciales entre tipos de clima diferentes. El ejemplo que traemos a colación representa los combos temporales de Santiago de Compostela (izquierda) y de Almería (derecha), los cuales representan al clima oceánico (Cfa) y el clima mediterráneo desértico (BWh). Tal y como se puede observar en la figura, resulta especialmente llamativa la diferencia de los combos situados en el vector W de la figura (precipitación) y E de la misma (frecuencia de días con lluvia), los cuales presentan un dimensionamiento diferencial que muestra la diferencia más importante entre ambos climas, húmedos (alta precipitación y alta frecuencia de lluvia el primero y baja precipitación y frecuencia de días lluviosos los segundos).

c) Definición de patrones climáticos mediante la herramienta de patrones recursivos (obsérvese la figura 7.7).

Esta herramienta gráfica es de gran utilidad para explorar los patrones climáticos relacionados con la estacionalidad del comportamiento de las variables climáticas. Gracias a la posibilidad de establecer patrones temporales diversos, la selección del patrón mensual y la distribución de los valores anuales, ofrecen la composición de una gráfica que representa perfectamente el comportamiento de cada una de las variables tratadas a lo largo de las distintas estaciones climáticas.

Aplicada esta técnica, en los puntos de estudio (obsérvese la figura 7.4) pueden observarse claramente los grandes patrones de variabilidad climática anual existentes en el conjunto de la península ibérica. Especialmente reseñables como ejemplo de la validez y utilidad de esta

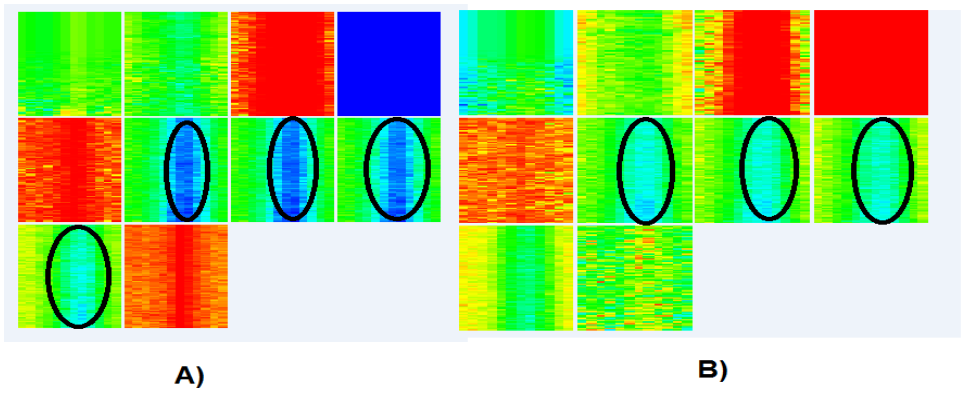

Figura 7.7 Patrones recursivos de las variables en Almería A) y San Sebastián B). Se ha utilizado una elipse para destacar los valores del verano. El orden de las variables es cld, dtr, frs, pet, pre, tmp, tmn, tmx, vap y wet, según muestra el gráfico de la siguiente figura para la zona de Santiago de Compostela.

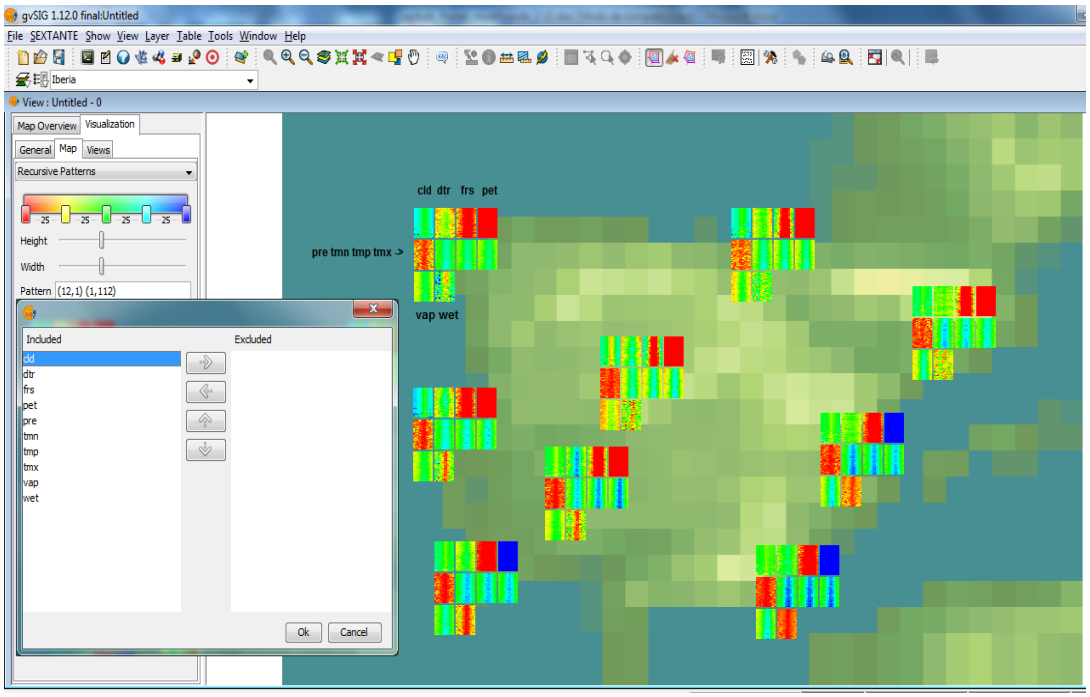

Figura 7.8 Patrones recursivos sobre los puntos característicos de la península ibérica.

técnica, son los claros comportamientos estacionales de los puntos más meridionales de la península ibérica, en los que el estío invernal (ausencia de nubosidad, pluviometría escasa y ausencia de días de lluvia), así como los altos valores termométricos quedan claramente marcados en las zonas centrales de las gráficas. En contraposición a estos comportamientos estivales, se encuentran las estaciones de la Iberia húmeda, específicamente los puntos de Santiago de Compostela y San Sebastián, marcadas por un estío mucho más húmedo y fresco derivado de su clima oceánico, que quedan también perfectamente reflejadas en la distribución de colores.

d) Definición de patrones climáticos mediante la herramienta de coordenadas de estrella.

Entre las herramientas gráficas diseñadas de mayor utilidad se tiene coordenadas de estrella, tanto por su estructura como por la configuración muy cercana a la ciencia climática (tal y como se ha mencionado anteriormente). Al ser muy similar a la configuración de las rosas de los

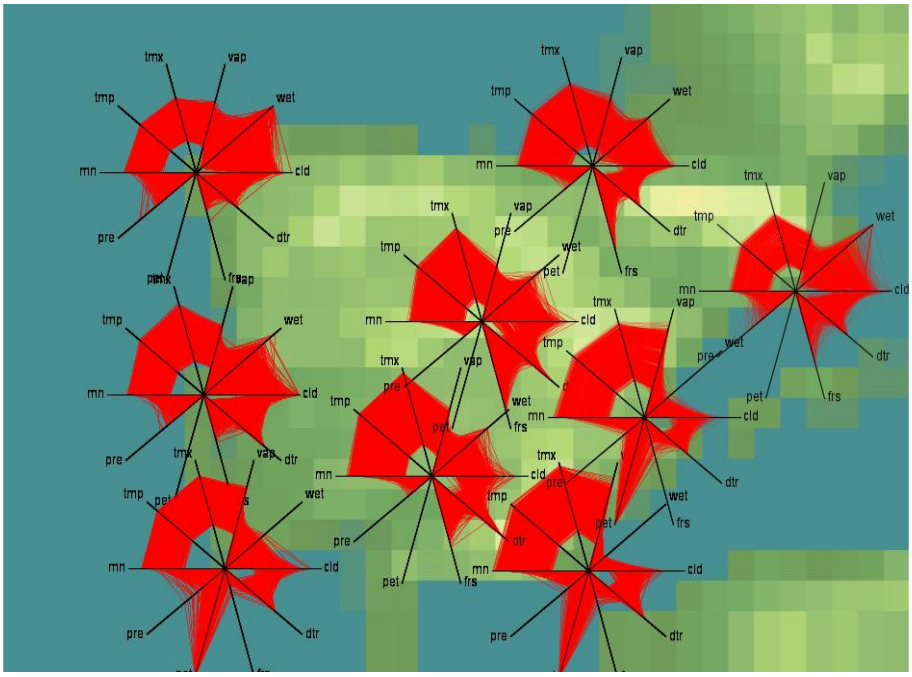

Figura 7.9 Coordenadas de estrella sobre los puntos característicos de la península ibérica.

vientos, muestra a partir de los distintos cuadrantes que configuran las distintas variables, una imagen muy gráfica de los gradientes y patrones climáticos de la península ibérica. Tal y como puede observarse en la figura 7.8, existe un claro gradiente NW-SE y N-S en la predominancia de los distintos cuadrantes de las figuras, en este caso el cuadrante NE (variables relacionadas con la frecuencia de días lluviosos) en los puntos septentrionales de la península ibérica y muy especialmente en el punto más noroccidental (Santiago de Compostela). Además, esta preponderancia se corresponde con una minoración del cuadrante NW de las figuras (variables relacionadas con la temperatura), por lo general, como término medio, en esta región más al norte se encuentran los climas más fríos. Este peso diferencial sufre una gradación inversa a lo largo de la península ibérica, y ocurre una transición de la preponderancia de esos dos cuadrantes hasta llegar a las latitudes meridionales de esta en la que la figura se invierte, y como resultado el sector NW de la figura es el más desarrollado (valores térmicos más altos), correlacionados con un sector NE de las figuras más debilitado como consecuencia de la disminución en los valores de precipitación y frecuencia de días lluviosos. Esta transición se corresponde perfectamente con la transición climática de la península ibérica que, de esta manera, queda muy explicitado de forma instantánea con la lectura de los gráficos de coordenadas de estrellas aplicadas a los puntos característicos.

e) Exploración de relaciones de variables climáticas mediante la herramienta gráfica de matriz de diagrama de dispersión.

En este apartado vamos a poner de manifiesto la utilidad evidente de esta técnica estadística genérica de visualización y análisis de las relaciones existentes entre distintos pares de variables. Por su naturaleza causal y las leyes físicas del funcionamiento del sistema climático, es

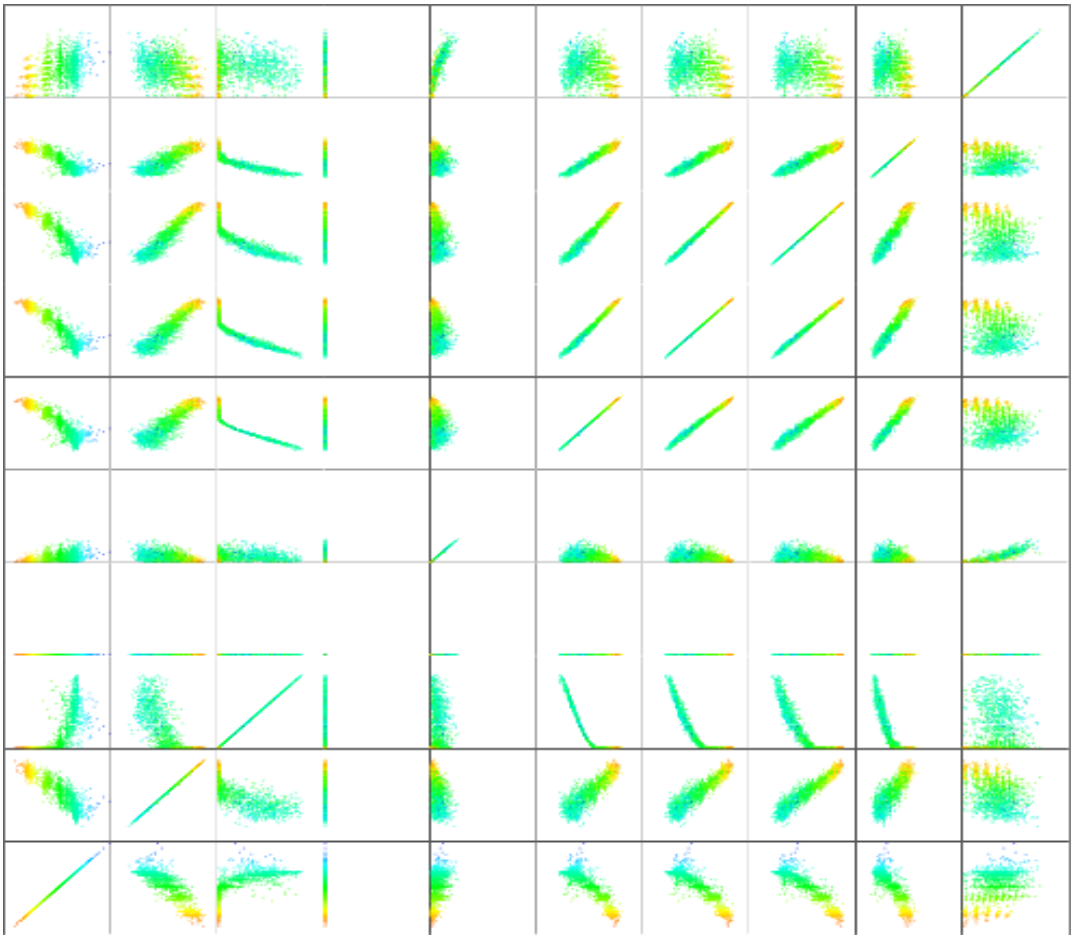

Figura 7.10 Matriz de diagramas de dispersión en la región de Madrid. Cada recuadro representa los valores de una variable contra otra. El orden de las variables es cld, dtr, frs, pet, pre, tmp, tmn, tmx, vap y wet, se comienza de abajo hacia arriba y de izquierda a derecha.

evidente que cuando manejamos información relativa a variables climáticas, la exploración y cuantificación del grado de asociación entre pares de variables que puedan mostrar un alto grado de correlación entre ellas, es una fuente de información de gran utilidad en la determinación y comparación de patrones climáticos entre áreas climáticas. Desde este punto de vista, podemos observar una serie de características diferenciales entre los diagramas de dispersión de distintas zonas climáticas que nos sirven para delimitar comportamientos diferenciales en los patrones climáticos y, por ende, para definir y clasificar a estas zonas en tipologías climáticas diferenciadas.

Observemos los casos de las figuras de los diagramas de dispersión entre variables de una zona de influencia atlántica como Lisboa en la costa oeste de la península ibérica y otra zona de marcado carácter continental como Madrid en la zona central de la misma (obsérvense las figuras 7.10 y 7.11.

Es evidente que existen relaciones entre variables marcadas por las leyes físicas fundamentales que gobiernan el comportamiento del sistema climático; por ejemplo entre cobertura nubosa y valores pluviométricos, o entre la temperatura mínima y el número de días de helada. Estas

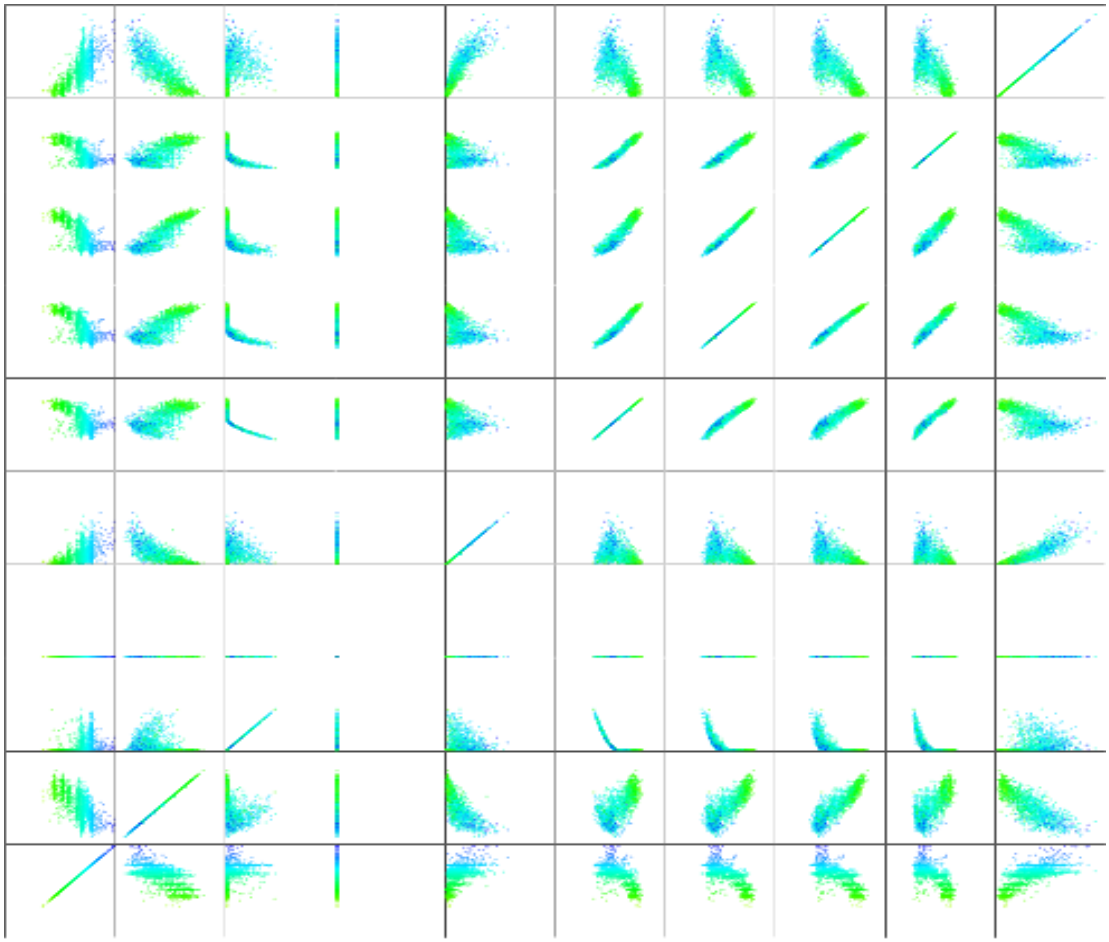

Figura 7.11 Matriz de diagramas de dispersión en la región de Lisboa. Cada recuadro representa los valores de una variable contra otra. El orden de las variables es cld, dtr, frs, pet, pre, tmp, tmn, tmx, vap y wet, de abajo hacia arriba y de izquierda a derecha.

relaciones son evidentemente persistentes entre las mismas con independencia del clima que estemos analizando. No obstante, es el grado o la fuerza de la asociación-correlación el que varía de unos climas a otros y el que va a permitir a partir del análisis visual de la herramienta, la matización entre distintos tipos de climas o zonas climáticas.

En el ejemplo previsto, esta disimetría oceanidad-continentalidad queda de manifiesto en algunas de esas asociaciones entre variables. Especialmente significativo desde este punto de vista es el diagrama de dispersión entre las variables frecuencia de días con lluvia y cobertura nubosa. En el primer caso, Lisboa, observamos que existe una correlación casi exponencial positiva muy fuerte entre los valores de ambas variables atmosféricas, lo que indica procesos generalizados de condensación atmosférica más intensos y de génesis de nubosidad de desarrollo vertical (*cumulonimbus*) que coinciden con los valores más altos en la frecuencia de días de lluvia en esa zona. La exposición de Lisboa a los vientos húmedos atlánticos del Oeste (*westerlies*), cargados de humedad y normalmente asociados a ascensos frontales del aire (*cold fronts*), implica en estas zonas un alto porcentaje de exposición a estos fenómenos de génesis de nubosidad de desarrollo vertical de origen frontal que están asociadas a los procesos de precipitación el aire.

Por el contrario, el alto grado de continentalidad asociado a la posición interior de Madrid en la península ibérica y la menor dependencia genética de los procesos de nubosidad respecto a la entrada de frentes fríos húmedos atlánticos, implican la existencia de un diagrama de dispersión en el que el grado de asociación de las dos variables es casi aleatorio, y muestra procesos claros de continentalidad en los cuales los procesos genéticos de las nubes y las tipologías de las mismas responden a mecanismos diferentes (nubosidad sin desarrollo vertical estratiforme y, por lo tanto, sin procesos de precipitación característicos de días fríos invernales o meses con mayor frecuencia de días de lluvia pero con poca cobertura nubosa en el mes), que son característicos de verano en los que los procesos de ascenso convectivos del aire estival extremadamente cálido (tormentas veraniegas), implican la existencia de varios días de precipitación en el mes, pero con presencia escasa de cobertura nubosa que aparece y desaparece con el fenómeno tormentoso.

7.2. Validación de los resultados

Para conocer la opinión de los expertos sobre la contribución de la integración de técnicas de visualización de datos multiparamétricos en sistemas de información geográfica, se empleó, como técnica de expertos, específicamente la técnica Delphi (Linstone y Turoff, 1975), que se basa en utilizar en la solución de problemas los juicios de un grupo de personas (expertos) con conocimientos teóricos y prácticos sobre la temática analizada, a través de un sistema de medición que permite ponderar aquellas apreciaciones cualitativas que se hayan realizado por estos expertos. El método de trabajo con expertos es utilizado actualmente en la informática y la ciencia de la computación, con buena aceptación (Obeso, 2005). En especialidades tales como ingeniería o administración, se entiende por calidad:

- El conjunto de atributos que poseen los bienes o servicios para satisfacer los requisitos de los consumidores.
- La totalidad de las características, de productos o servicios, que soportan su capacidad para satisfacer las necesidades sugeridas o establecidas.

Según Meservy *et al.* (2012) la calidad del *software* es:

- El grado en el que el software satisface una serie de requisitos de operación preestablecidos, los estándares de desarrollo especificados con anterioridad y las características inherentes a todo producto de software desarrollado de manera profesional.
- La eficiencia, efectividad y facilidad de uso del software, bajo determinadas condiciones, para un conjunto de usuarios con propósitos específicos.

La opinión de los expertos sobre la contribución de la propuesta, se conoce aplicando la escala psicométrica creada por Likert (1932).

Se definieron nueve planteamientos para valorar los aspectos siguientes:

- La utilidad de visualizar múltiples variables temporales a la vez para extraer relaciones, patrones, tendencias y anomalías que están presentes en los datos.
- La interpretación de los datos una vez comprendida una técnica de visualización.
- La facilidad para asimilar los conocimientos mínimos que permitan manipular formatos de datos científicos en sistemas de información geográfica.
- La facilidad para comprender la estructura y comportamiento de grandes conjuntos de datos realizando análisis exploratorio de datos con técnicas de visualización científica.
- La facilidad que brindan los algoritmos para la manipulación de formatos de datos científicos en sistemas de información geográfica para la preparación de conjuntos de datos que pueden ser analizados por herramientas de visualización.
- El nivel de interacción que poseen las técnicas de visualización del módulo de visualización científica desarrollado para gvSIG.
- La facilidad de la interfaz del sistema para el uso y manipulación de las técnicas de visualización.
- El desempeño del sistema propuesto para el análisis de grandes volúmenes de datos.
- La utilidad de la integración de visualización de datos multiparamétricos en sistemas de información geográfica para la interpretación simultánea de múltiples variables.

Para la evaluación se diseñó una encuesta (obsérvese anexo 4), la cual formula una pregunta para cada planteamiento. Este instrumento fue aplicado a especialistas con conocimientos y/o experiencias en sistemas de información geográfica y visualización científica. Los expertos podían seleccionar para cada pregunta una respuesta de las siguientes, que se interpretarán con el valor escrito entre paréntesis: Muy Alto (5), Alto (4), Neutro (3), Bajo (2) y Muy Bajo (1).

7.2.1. Resultados de la encuesta de selección de expertos

Para una valoración inicial de los posibles expertos, fueron contactados profesionales que actualmente trabajan en Cuba, España, Portugal y Francia. Se consultaron profesores e investigadores españoles de la Universidad de Sevilla y de la Asociación gvSIG. Se cuenta con los criterios como experto de al menos un investigador de la Universidad de las Azores, en Portugal y otro del Instituto de Investigaciones en Informática y Sistemas Aleatorios (IRISA, por sus siglas en francés) en Rennes, Francia. De igual manera, se contó con la ayuda de especialistas cubanos: del Centro de Estudios y Servicios Ambientales de la provincia de Villa Clara (CESAM VC), Instituto de Meteorología Provincial de Villa Clara, la empresa de Cartografía y Soluciones Geomáticas GeoSI, la Empresa de Investigaciones y Proyectos Hidráulicos de Villa Clara, del Instituto Superior Politécnico José Antonio Echeverría (ISPJAE), y de profesores de Física y Computación de la Universidad Central “Marta Abreu” de Las Villas (UCLV).

Tabla 7.11 Fuentes de argumentación de los expertos.

No	Fuentes de argumentación	Alto (A)	Medio (M)	Bajo (B)
1	Estudios teóricos realizados por usted.	0,30	0,2	0,10
2	Experiencia adquirida durante su vida profesional.	0,50	0,37	0,30
3	Conocimiento de investigaciones y/o publicaciones nacionales e internacionales.	0,05	0,04	0,03
4	Conocimiento propio sobre el estado del tema de investigación.	0,05	0,04	0,03
5	Actualización en cursos de postgrado, diplomados, maestrías, doctorado, etc.	0,05	0,04	0,03
6	Intuición.	0,05	0,03	0,02

Para asegurar la confiabilidad de las respuestas, se evaluó la idoneidad de los expertos en esta temática (obsérvese anexo 5), mediante el cálculo de su coeficiente de competencia.

Para determinar el coeficiente de competencia de los candidatos a expertos se aplicó el cálculo de dicho coeficiente de la forma siguiente:

$$Kcomp = \frac{1}{2} * (K_c + K_a)$$

donde:

- $Kcomp$: Coeficiente de competencia.
- K_c: Coeficiente de conocimiento o información que tiene el experto acerca del problema, calculado sobre la valoración del propio experto en una escala de 0 a 10 y multiplicado por $0,1$.
- K_a: Coeficiente de argumentación o fundamentación de los criterios del experto, obtenido como resultado de la suma de los puntos de acuerdo con la tabla patrón 7.11:

Producto del análisis de las encuestas complementarias aplicadas a los expertos acerca del conocimiento del tema de investigación se pudo comprobar que el **100** por ciento de los encuestados presentó un nivel de competencia alta y media, por lo cual todos estos fueron seleccionados para el análisis con las encuestas de validación de las herramientas. La tabla 7.12 muestra los resultados obtenidos luego de asignar los pesos correspondientes para cada una de las categorías.

La composición de los expertos teniendo en cuenta su categoría científica y su cargo, se refleja en la tabla 7.13.

La tabla 7.14 muestra la cantidad de investigadores por instituciones.

Tabla 7.12 Nivel de competencia de los expertos.

No	EA	ET	I	CP	A	I	K_a	K_c	$(K_a + K_c)/2$	Competencia
1	0,37	0,2	0,04	10	0,03	0,03	0,67	1	0,835	Alta
2	0,5	0,1	0,03	5	0,03	0,05	0,71	0,5	0,605	Media
3	0,5	0,1	0,04	7	0,03	0,03	0,7	0,7	0,7	Media
4	0,3	0,1	0,03	6	0,03	0,03	0,49	0,6	0,545	Media
5	0,5	0,3	0,05	10	0,05	0,03	0,93	1	0,965	Alta
6	0,37	0,2	0,04	7	0,05	0,05	0,71	0,7	0,705	Alta
7	0,37	0,2	0,04	10	0,03	0,05	0,69	1	0,845	Alta
8	0,5	0,3	0,04	5	0,05	0,05	0,94	0,5	0,72	Alta
9	0,37	0,2	0,04	5	0,03	0,05	0,69	0,5	0,595	Media
10	0,37	0,2	0,05	8	0,03	0,05	0,7	0,8	0,75	Alta
11	0,37	0,2	0,04	5	0,03	0,03	0,67	0,5	0,585	Media
12	0,5	0,2	0,04	10	0,03	0,05	0,82	1	0,91	Alta
13	0,37	0,2	0,04	5	0,04	0,03	0,68	0,5	0,59	Media
14	0,5	0,3	0,03	7	0,03	0,05	0,91	0,7	0,805	Alta
15	0,37	0,2	0,04	8	0,04	0,05	0,7	0,8	0,75	Alta
16	0,37	0,2	0,04	5	0,05	0,03	0,62	0,5	0,595	Media
17	0,37	0,2	0,04	5	0,03	0,02	0,66	0,5	0,58	Media
18	0,37	0,2	0,04	7	0,03	0,03	0,67	0,7	0,685	Media
19	0,37	0,2	0,04	6	0,04	0,05	0,67	0,6	0,65	Media
20	0,37	0,2	0,04	8	0,04	0,05	0,7	0,8	0,75	Alta
21	0,37	0,2	0,04	4	0,04	0,03	0,68	0,4	0,54	Media
22	0,37	0,2	0,04	7	0,04	0,03	0,68	0,7	0,69	Media
23	0,37	0,1	0,04	8	0,04	0,05	0,6	0,8	0,7	Media
24	0,5	0,1	0,04	6	0,04	0,05	0,73	0,6	0,665	Media
25	0,37	0,2	0,04	5	0,04	0,03	0,68	0,5	0,59	Media

Tabla 7.13 Composición de los expertos involucrados en la validación.

Perfil de trabajo	Cantidad	%		Nivel académico	Cantidad	%
Profesor	13	52		Doctores	13	52
Investigador	6	24		Máster	10	40
Empresario	6	24		No especificado	2	8

Tabla 7.14 Composición por instituciones.

Institución	**Cantidad**	**%**
Universidad Central de Las Villas	4	16
Universidad de Sevilla	6	24
Instituto de Meteorología de Villa Clara	2	8
Universidad de las Azores	1	4
Empresa de Cartografía y Soluciones Geomáticas GeoSI	5	25
Centro de Estudios y Servicios Ambientales de Villa Clara	2	8
Instituto Superior Politécnico José Antonio Echeverría	1	4
Empresa de Investigaciones y Proyectos Hidráulicos de Villa Clara	1	4
Instituto de Investigaciones en Informática y Sistemas Aleatorios de Rennes	1	4
Asociación gvSIG	2	8

7.2.2. Resultados de la encuesta de validación de los métodos y herramientas

De los **25** expertos, **10** y **15** de ellos obtuvieron un coeficiente de competencia alto y medio, respectivamente. Cada uno de ellos realizó la encuesta que aparece en el anexo 4. La encuesta aplicada se fundamentó en los valores de la escala de Likert, se calcularon los por cientos de concordancia de los expertos con cada una de las posibles respuestas para los planteamientos formulados. Luego se calcula un índice porcentual (IP) que integra en un solo valor la aceptación de cada planteamiento por los evaluadores mediante la fórmula siguiente:

$$IP = \frac{5 * (\%MA) + 4 * (\%A) + 3 * (\%Neutro) + 2 * (\%B) + 1 * (\%MB)}{5}$$

En el gráfico de la figura 7.12, se puede apreciar cómo se comporta el índice porcentual para cada una de las variables que evalúan las opiniones de los expertos acerca de la utilidad de integrar técnicas de visualización científica en sistemas de información geográfica para el almacenamiento y análisis de grandes volúmenes de datos espacio-temporales. Cada una de las preguntas de la encuesta se etiqueta con una letra "u", seguida del número de la encuesta que varía desde 1 hasta 9.

Los índices porcentuales para todas las preguntas sobrepasa el valor de 75, siendo superior a 90 para las preguntas u1 y u9. El resto de las preguntas oscila entre los valores 75 y 90.

Los mejores resultados se obtuvieron para las preguntas u1 y u9, por lo que se puede decir que los expertos de conjunto consideran que es de gran utilidad visualizar múltiples variables temporales a la vez para extraer relaciones, patrones, tendencias y anomalías presentes en los datos, y que esto se puede llevar a cabo mediante la integración de visualización de datos multiparamétricos en sistemas de información geográfica para la interpretación simultánea de múltiples variables. Ambas son ideas que se han desarrollado con éxito en esta tesis.

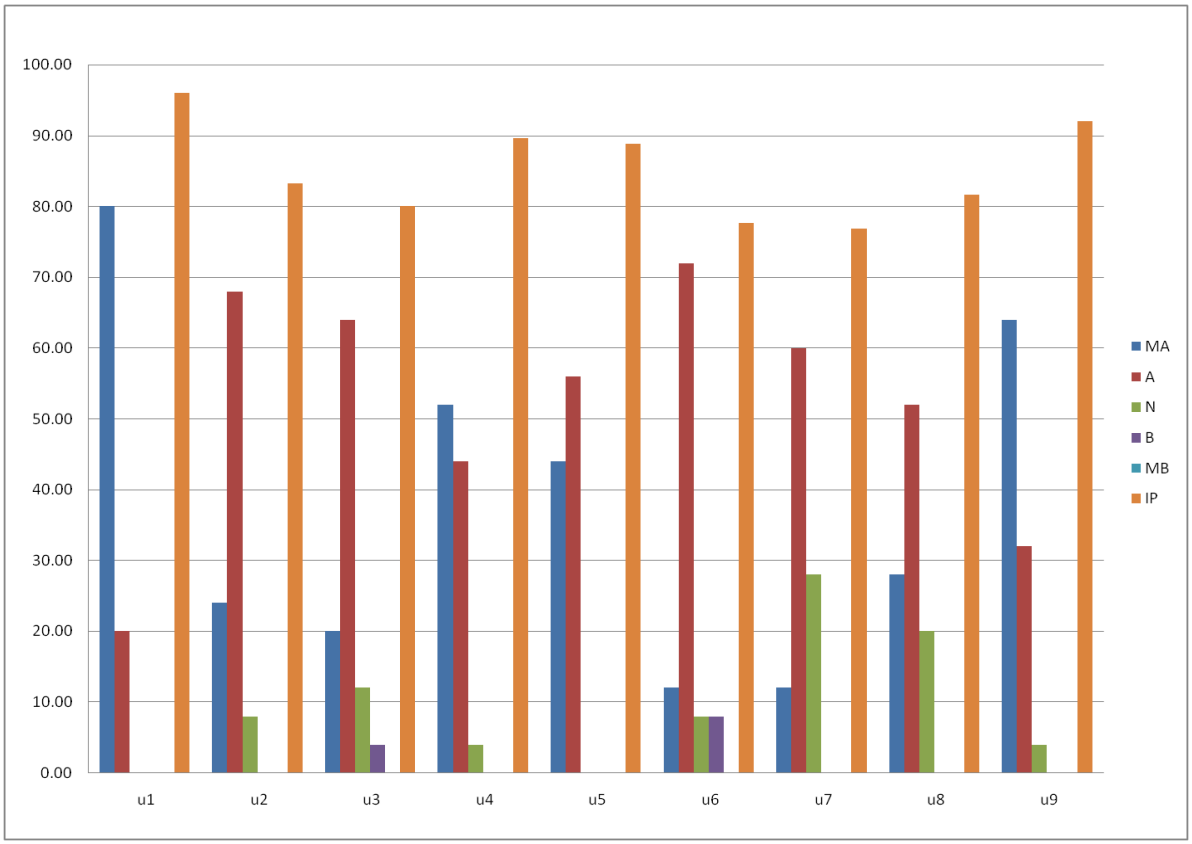

Figura 7.12 Valoración de los expertos sobre la utilidad de la integración de técnicas de visualización científica en sistemas de información geográfica para el almacenamiento y análisis de grandes volúmenes de datos espacio-temporales. Se representa con barras de diferentes colores, la cantidad de expertos que respondieron Muy Alto (MA), Alto (A), Neutro (N), Bajo (B), Muy Bajo (MB) y el índice porcentual.

Se considera un resultado bueno, la opinión de los expertos con respecto a: la interpretación de los datos una vez comprendida una técnica de visualización, la facilidad para asimilar los conocimientos mínimos que permitan manipular formatos de datos científicos en sistemas de información geográfica, y la facilidad para comprender la estructura y comportamiento de grandes conjuntos de datos realizando análisis exploratorio de datos con técnicas de visualización científica. Algunos de los expertos valoraron como neutro o bajo los aspectos antes mencionados. De igual manera, se considera como buenos resultados, la facilidad que brindan los algoritmos para la manipulación de formatos de datos científicos en sistemas de información geográfica, para la preparación de conjuntos de datos, que pueden ser analizados por herramientas de visualización. Se debe prestar especial atención en el desempeño de los sistemas propuestos para el análisis de grandes volúmenes de datos.

Los temas con índices porcentuales más cercanos a 75, se pueden considerar los resultados menos favorables. Estos son aspectos en los que se debe seguir trabajando para mejorarlos, y tienen que ver con el nivel de interacción que poseen las técnicas de visualización de las herramientas desarrolladas y las interfaces para el uso y la manipulación de las técnicas de visualización en sistemas de información geográfica.

7.3. Conclusiones parciales

Mediante los casos de estudios presentados en este capítulo se ha comprobado la utilidad de los métodos y herramientas desarrollados para el análisis visual de grandes volúmenes de datos climáticos. Se estudiaron los tipos de clima de la península ibérica y se mostraron ejemplos que evidencian los contrastes, y resaltan propiedades poco estudiadas sobre los subtipos de clima de la península.

En este capítulo se aplicó el método de expertos para realizar una validación de las propuestas realizadas en esta tesis. En particular se obtuvo una valoración sobre la contribución de la integración de técnicas de visualización de datos multiparamétricos en sistemas de información geográfica para el análisis exploratorio de grandes volúmenes de datos científicos.

Se calculó la competencia de cada uno de los expertos, basado en sus propios criterios respecto a conocimientos teóricos sobre este tema y la experiencia acumulada en su vida profesional. Se aplicaron dos encuestas a los expertos seleccionados y se pudo obtener una valoración general del conjunto de expertos acorde a los aportes presentados en este trabajo. En el procesamiento de las encuestas aplicadas se obtuvieron excelentes resultados sobre la investigación realizada y se evidenciaron conclusiones que ratifican la importancia de esta investigación; de igual manera se resaltaron aspectos en los que se debe seguir trabajando.

Parte III

Conclusiones

8 Conclusiones y trabajos futuros

En este capítulo se hace un resumen de las principales conclusiones de este trabajo y se describen un conjunto de temas abiertos que han surgido en el transcurso de esta investigación, que merecen ser estudiados en trabajos futuros.

8.1. Conclusiones

En este trabajo se ha dado solución a los problemas que originaron esta investigación mediante el cumplimiento del objetivo general trazado en la sección 1.2. Se ha creado un modelo que permite integrar herramientas de visualización en sistemas de información geográfica para el análisis de grandes volúmenes de datos espacio-temporales, brindando la posibilidad de analizar simultáneamente múltiples variables geo-referenciadas, que pueden representar series temporales.

Del estudio sistemático de las técnicas de visualización de datos multiparamétricos y su modificación para la integración con sistemas de información geográfica, que permitan el análisis exploratorio de grandes volúmenes de datos, se seleccionaron conjuntos de técnicas de visualización que son adecuadas para esta integración. El estudio de las principales características de algunos de los sistemas de información geográfica más populares permitió hacer una valoración, que finalmente resultó en la selección de gvSIG y Sextante como herramientas para la incorporación de los conceptos y modelos desarrollados en esta tesis. Luego de estudiar algunas de las principales herramientas que permiten realizar análisis exploratorio de datos mediante la integración de técnicas de visualización y el uso de mapas y formatos de sistemas de información geográfica, se pudieron analizar sus principales ventajas y deficiencias. Este estudio posibilitó realizar valoraciones sobre elementos novedosos en los que se trabajó, como la manipulación de grandes volúmenes de datos científicos y las interacciones y modificaciones realizadas a las técnicas de visualización seleccionadas para su integración en sistemas de información geográfica. Los formatos de datos científicos seleccionados para la integración con las herramientas de visualización en sistemas de información geográfica fueron HDF y netCDF.

En esta tesis los principales algoritmos y herramientas que hacen uso de estos formatos de datos científicos fueron descritos en los capítulos 5 y 6.

Se desarrolló un modelo que permitió definir las pautas para integrar técnicas de visualización en sistemas de información geográfica facilitando la manipulación e integración de datos geográficos y multiparamétricos. Este modelo se llevó a cabo mediante el desarrollo de herramientas para el análisis exploratorio de datos con baja y alta densidad espacial, incorporando en este último caso la manipulación de formatos de datos científicos. En los capítulos 4, 5 y 6 se mostraron los principales resultados de la aplicación del modelo propuesto. El modelo puede ser generalizado y aplicado con herramientas de visualización, sistemas de información geográfica y formatos de datos científicos similares o diferentes a los seleccionados en esta tesis. Las principales ideas del modelo propuesto fueron publicadas en la *Revista Técnica de la Facultad de Ingeniería de la Universidad del Zulia* (Vázquez-Rodríguez *et al.*, 2015).

El análisis exploratorio de datos con baja densidad espacial fue tratado en el capítulo 4 como un caso especial que permite analizar visualmente datos multiparamétricos de manera independiente o coordinada asociados a pocos lugares del espacio. Se obtuvo una herramienta con múltiples técnicas de visualización de datos multiparamétricos que fueron integradas como una extensión de la versión 1.9 de gvSIG. Se presentó un caso de estudio con datos climáticos de la provincia de Villa Clara, Cuba y los principales resultados fueron publicados en la revista brasileña *Anuario do Instituto de Geociências* (Vázquez-Rodríguez *et al.*, 2013b) y presentados en un evento internacional de prestigio (Vázquez-Rodríguez *et al.*, 2010a). Estas herramientas pueden ser generalizadas y utilizadas en múltiples áreas de aplicación debido a que son generales y pueden ser configuradas para múltiples propósitos dentro del análisis exploratorio de datos.

En el capítulo 6 se presentó una solución a los problemas del análisis exploratorio de datos con alta densidad espacial. Para eso se aplicó el modelo propuesto en el capítulo 3 y se obtuvo una herramienta que permite la extracción de conocimiento, reconocimiento de patrones, tendencias, anomalías y relaciones entre múltiples variables distribuidas uniformemente sobre un territorio, estas variables pueden ser series temporales y por la alta densidad espacial de que se dispone, los volúmenes de información con los que se trabaja pueden llegar a ser difíciles de manipular. Se propuso la integración de un conjunto de técnicas de visualización de datos multiparamétricos en la versión 1,12 de gvSIG. Estos resultados, igualmente fueron publicados en (Vázquez-Rodríguez *et al.*, 2015).

Un conjunto de herramientas y algoritmos que permiten la manipulación de los formatos de datos científicos HDF y netCDF en sistemas de información geográfica fueron descritos en capítulo 6. Las herramientas y algoritmos fueron implementados como una extensión de la biblioteca Sextante, por lo que pueden ser utilizadas en cualquier sistema de información geográfica basado en Java que permita la integración de esta biblioteca. La extensión desarrollada en Sextante es de gran utilidad para la transformación de formatos de datos comunes en los sistemas de información geográfica para formatos de datos científicos y viceversa. Además, algunos de los algoritmos están diseñados para facilitar la creación automática de conjuntos de

datos con la estructura necesaria que requiere el módulo de visualización científica de gvSIG descrito en el capítulo 6. Se demostró el uso de las herramientas desarrolladas mediante un caso de estudio que evidencia su utilidad para crear grandes conjuntos de datos que pueden ser analizados mediante visualizaciones en sistemas de información geográfica.

Ha quedado comprobada mediante los casos de estudio mostrados en esta tesis, la viabilidad de la utilización del modelo para la extracción de conocimiento, reconocimiento de patrones, tendencias, anomalías y relaciones entre múltiples variables en diferentes áreas de aplicación, así como la formulación y corroboración de hipótesis. En la sección 7.2 se aplicó el método de expertos para realizar una validación de las propuestas realizadas en esta tesis. En particular se obtuvo una valoración sobre la contribución de la integración de técnicas de visualización de datos multiparamétricos en sistemas de información geográfica para el análisis exploratorio de grandes volúmenes de datos científicos. Se calculó la competencia de cada uno de los expertos, basado en sus propios criterios sobre conocimientos teóricos sobre este tema y la experiencia acumulada en su vida profesional. Se aplicaron dos encuestas a los expertos seleccionados y se pudo obtener una valoración general del conjunto de expertos sobre los aportes presentados en este trabajo. En el procesamiento de las encuestas aplicadas se obtuvieron excelentes resultados sobre la investigación realizada y se evidenciaron conclusiones que ratifican la importancia de esta investigación. De igual manera se resaltaron aspectos en los que se debe seguir trabajando, los cuales se presentan en la siguiente sección.

8.2. Trabajos futuros

En el desarrollo de esta investigación y teniendo en cuenta los principales resultados obtenidos, ha surgido un conjunto de ideas y motivaciones que, más allá del alcance de esta tesis, se pueden presentar como temas para desarrollar en trabajos futuros.

En primer lugar, se debe seguir explotando y enriqueciendo el modelo propuesto para continuar obteniendo herramientas para el análisis exploratorio de grandes volúmenes de datos en otros contextos como es el caso de Big Data. En particular, se deben crear nuevas herramientas de visualización con niveles de interacción amigables, y así permitir que usuarios de varias ramas de la ciencia puedan utilizar otras interfaces para el uso y manipulación de grandes volúmenes de datos, como es el caso de las nuevas tendencias de la interacción humano-máquinas (HCI por sus siglas en ingles de *Human Computer Interaction*). Se debe prestar especial atención en el desempeño de los nuevos sistemas que se creen para análisis de grandes volúmenes de datos, para esto se puede trabajar con otros formatos de datos científicos y otras técnicas que no se hayan utilizado en este trabajo. Los métodos y herramientas desarrollados en esta tesis han sido utilizados en campos de aplicación como la meteorología y la climatología; sin embargo, estos son generales y se deben utilizar en otros campos de aplicación.

8.3. Publicaciones derivadas de la investigación

Durante el desarrollo de la presente investigación se obtuvieron los resultados que se muestran a continuación.

Artículos en revistas:

- Vázquez-Rodríguez, Romel; Pérez-Risquet, Carlos; y Torres, Juan Carlos. 2015. Exploratory data analysis through the integration of visualization techniques in geographical information systems, *Revista Técnica de la Facultad de Ingeniería de la Universidad del Zulia*, 38(1), 73–82, ISSN:0254-0770, factor de impacto (2014): 0.047, SJR: 0.12.
- Vázquez-Rodríguez, Romel; Pérez-Risquet, Carlos; y Torres, Juan Carlos. 2013. A novel visual data mining module for the geographical information system gvsig, *Anuário do Instituto de Geociências*, 36(1), 98–111, ISSN:0101-9759, SJR: 0.25.

Trabajos presentados en congresos por el autor de la tesis:

- Uso de sistemas de información geográfica basados en software libre para la visualización de datos meteorológicos. Informática 2009. Cuba. ISBN: 978-959-286-010-0.
- Estudio de la factibilidad de la aplicación de técnicas de visualización de datos multiparamétricos para el análisis visual de datos meteorológicos. CompuMat 2009. Cuba. ISBN: 1728-6042.
- Visualización de datos meteorológicos mediante técnicas de visualización científica en SIG. Seminario Internacional de Doctorado en Soft Computing. Cuba. ISBN: 959-250-525-4.
- Implementation of scientific visualization techniques for meteorological visual data in GIS. Cuba-Flanders Workshop on Machine Learning and Knowledge Discovering CFWMLKD 2010. Cuba. ISBN: 578959250574-2.
- A new visual data mining tool for gvSIG GIS. International Conference on Knowledge Discovery and Information Retrieval. KDIR 2010. España. ISBN: 978-989-8425-28-7.
- Herramientas de visualización para el análisis visual de datos espacio-temporales en gvSIG. Informática 2011. Cuba. ISBN: 978-959-7213-01-7.
- Visualización Espacio-Temporal mediante técnicas de visualización de datos multiparamétricos en SIG. Geociencias 2011. Cuba. ISBN: 978-959-7117-30-8
- Extensión del módulo de visualización de datos espacio-temporales de gvSIG. Informática 2013. Cuba. ISBN: 978-959-7213-02-4.
- A framework for the visualization and analysis of environmental and climatic temporal sequences evenly distributed in space and time. Workshop and Annual Meeting of the DAAD funded Network “Developing Sustainability”. 2013. Indonesia.
- Un nuevo método para el análisis visual de grandes volúmenes de datos espacio-temporales. Geociencias 2015. Cuba.

- Algoritmos para la manipulación de formatos de datos científicos en gvSIG. Geociencias 2015. Cuba.

Premios y condecoraciones obtenidas por el autor de la tesis durante la formación doctoral:

- Premio Anual Provincial de la Academia de Ciencias de Cuba. 2011. Herramientas y técnicas para la visualización científica en sistemas de información geográfica.
- Título de experto universitario en sistemas de información geográfica, otorgado por la Universidad Internacional de Andalucía. 2012.
- Premio Anual Provincial de la Academia de Ciencias de Cuba. 2013. Integración de técnicas de visualización de datos en sistemas de información geográfica.
- Premio Nacional 2014 “Gustavo Furrazola Bermúdez”. Otorgado por la Sociedad Cubana de Geología a jóvenes destacados en la investigación en las Geociencias.

Bibliografía

Abello, J. y J. Korn. 2002. Mgv: A system to visualize massive multi-digraphs, Citeseer.

AEMET, IM. 2011. Atlas climático ibérico/ iberian climate atlas, *Agencia Estatal de Meteorología, Ministerio de Medio Ambiente y Rural y Marino, Madrid*.

Aigner, Wolfgang. 2011. *Visualization of time-oriented data*, Springer.

Aigner, Wolfgang, Silvia Miksch, Wolfgang Muller, Heidrun Schumann, y Christian Tominski. 2008. Visual methods for analyzing time-oriented data, *Visualization and Computer Graphics, IEEE Transactions on*, 14(1), 47–60.

Andrews, D. F. 1972. Plots of high dimensional data, *Biometric*, 28(1), 125–137, URL http://www.jstor.org/stable/2528964.

Andrews, Keith. 2005. Information visualisation, URL http://courses.iicm.edu/ivis/.

Andrienko, Gennady, Natalia Andrienko, Sebastian Bremm, Tobias Schreck, Tatiana Von Landesberger, Peter Bak, y Daniel Keim. 2010a. Space-in-time and time-in-space self-organizing maps for exploring spatiotemporal patterns, *Computer Graphics Forum*, 29(3), 913–922.

Andrienko, Gennady, Natalia Andrienko, Urska Demsar, Doris Dransch, Jason Dykes, Sara Irina Fabrikant, Mikael Jern, Menno-Jan Kraak, Heidrun Schumann, y Christian Tominski. 2010b. Space, time and visual analytics, *International Journal of Geographical Information Science*, 24(10), 1577–1600.

Andrienko, N, G Andrienko, y P Gatalsky. 2003. Exploratory spatio-temporal visualization: an analytical review, *Journal of Visual Languages and Computing*, 14(6), 503–541, URL http://www.sciencedirect.com/science/article/B6WMM-49H1102-1/2/e0ae4bae5f05168db1b15cabce525e36.

Andrienko, Natalia y Gennady Andrienko. 2006. *Exploratory analysis of spatial and temporal data*, Berlin, Germany: Springer-Verlag.

Anguix, Álvaro. 2009. gvsig: un proyecto global casos de éxito.

Anguix, Álvaro y Gabriel Carrión. 2005. gvsig: Soluciones open source en las tecnologías espaciales, en *GISPLANET 2005*.

Ankerst, M., D. A. Keim, y H. P. Kriegel. 1996. 'circle segments': A technique for visually exploring large multidimensional dataset, en *Visualization*, San Francisco.

Beddow, J. 1990. Shape coding of multidimensional data on a microcomputer display, en *1990 IEEE conference on visualization*, Los Alamitos, CA.: IEEE Computer Society Press, 238–246.

Bertin, Jacques. 1983. *Semiology of graphics: diagrams, networks, maps*, University of Wisconsin Press, Madison.

Bolstad, P. 2005. *GIS fundamentals: A first text on geographic information systems*, Eider Pr.

Borrell González, Alejandro. 2012. Aplicación informática para la integración de los datos generados por el observatorio obsea en las redes de sistemas de observación.

Brio, Bonifacio Martin del y Alfredo Sanz Molina. 2001. *Redes neuronales y sistemas borrosos*.

Burns, R. y A. Skupin. 2013. Towards qualitative geovisual analytics: A case study involving places, people, and mediated experience, *Cartographica*, 48(3), 157–176, URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84884570196&partnerID=40&md5=d241d5f9bb364c86ffb167e8b482a5ec.

Burt, J.E. y G.M. Barber. 1996. *Elementary Statistics for Geographers, 2nd edn*, Guilford, New York.

Castro, R., J. Vega, M. Ruiz, D. Sanz, y E. Barrera. 2013. Analysis of netcdf-4 and hdf-5 scientific file formats for the data archiving of an iter fast plant system controller prototype, en *19th IMEKO TC4 Symposium - Measurements of Electrical Quantities 2013 and 17th International Workshop on ADC and DAC Modelling and Testing*, 699–704, URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84894206062&partnerID=40&md5=aab01a74b590a9fd8b24745e8ab6fa8f.

Chen, Chun-houh, Wolfgang Härdle, y Antony Unwin, (Eds.) . 2008. *Handbook of Data Visualization*, Berlin Heidelberg, Berlin: Springer-Verlag.

Chernoff, H. 1973. The use of faces to represent points in k-dimensional space graphically, *Journal of the American Statistical Association*, 68, 361–368.

Chorley, Lord Roger. 1987. *Handling geographic information: report to the secretary of state for the environment of the committee of enquiry into the handling of geographic information*, vol. 1, HM Stationery Office.

Clarke, Keith. 1990. *Analytical and Computer Cartography*, Prentice Hall Professional Technical Reference, first edn.

Cleveland, W.S. 1993. Visualizing data, en *Hobart Press*, Summit New Jersey.

Cook, Dianne, Jürgen Symanzik, James J. Majurea, y Noel Cressieb. 1997. Dynamic graphics in a gis: more examples using linked software, *Computers & Geosciences*, 23, 371–385.

Cowen, David. 1989. Ncgia lecture.

Cui, Q, MO Ward, y EA Rundensteiner. 2006. Enhancing scatterplot matrices for data with ordering or spatial attributes, en *SPIE*, Citeseer, vol. 6060, 248–258.

Dykes, J, AM MacEachren, y MJ Kraak. 2005. Exploring geovisualization, chapter 1, en *Exploring geovisualization*, Amsterdam: Elsevier, cap. Chapter 1, 3.

Eick, SG. 2000. Visualizing multi-dimensional data, *ACM SIGGRAPH computer graphics*, 34(1), 61–67.

Esparza Gil, Josefa. 2014. Contraste espacio-temporal de indicadores de interés hidrológico derivados desde teledetección.

Furnas, G.W. y A. Buja. 1994. Prosection views: dimensional inference through sections and projections, *Journal of Computational and Graphical Statistics*, 3(4), 323–353.

Gahegan, Mark. 2005. Exploring geovisualization, chapter 4. beyond tools: Visual support for the entire process of giscience, en *Exploring geovisualization*, Amsterdam: Elsevier, cap. Chapter 4.

Gahegan, Mark, Masahiro Takatsuka, Mike Wheeler, y Frank Hardisty. 2002. Introducing geovista studio: an integrated suite of visualization and computational methods for exploration and knowledge construction in geography, *Comput Environ Urban*, 26(4), 267–292.

GRASS. 2008. Geographic resources analysis support system, URL http://grass.itc.it.

Gray, James. 2008. Quantum gis: the open-source geographic information system, *Linux Journal*, 2008(172), 8.

Gray, Jim, David T Liu, Maria Nieto-Santisteban, Alex Szalay, David J DeWitt, y Gerd Heber. 2005. Scientific data management in the coming decade, *ACM SIGMOD Record*, 34(4), 34–41.

Group, FITS Working. 2009. Definition of the flexible image transport system (fits), *FITS Standard Version*, 3.

Group, HDF. 2011. Hdf5 user's guide.

Guo, D, J Chen, AM MacEachren, y K Liao. 2006. A visualization system for space-time and multivariate patterns (vis-stamp), *IEEE transactions on Visualization and Computer Graphics*, 12(6), 1461–1474.

Hand, D.J. 1999. Intelligent data analysis: an introduction, Springer, Berlin, Heidelberg 1999.

Hanisch, RJ, A Farris, EW Greisen, WD Pence, BM Schlesinger, PJ Teuben, RW Thompson, y A Warnock. 2000. Definition of the flexible image transport system (fits), *Astronomy and Astrophysics*.

Hankin, Steve C, Jon D Blower, Thierry Carval, Kenneth S Casey, Craig Donlon, Olivier Lauret, Thomas Loubrieu, Ashwanth Srinivasan, Joaquim Trinanes, y Oystein Godoy. 2010. Netcdf-cf-opendap: Standards for ocean data interoperability and object lessons for community data standards processes, en *Oceanobs 2009, Venice Convention Centre, 21-25 septembre 2009, Venise*.

Hansen, Charles D, Min Chen, Christopher R Johnson, Arie E Kaufman, y Hans Hagen. 2014. *Scientific Visualization: Uncertainty, Multifield, Biomedical, and Scalable Visualization*, Springer.

Hansen, Charles D. y Chris R. Johnson. 2005. *The visualization handbook*, Elsevier.

Harris, I, PD Jones, TJ Osborn, y DH Lister. 2014. Updated high resolution grids of monthly climatic observations - the cru ts3. 10 dataset, *International Journal of Climatology*, 34(3), 623–642.

Havre, S., B. Hetzler, y L. Nowell. 2000. Themeriver: Visualizing theme changes over time, IEEE, 115–123.

Hearnshaw, Hilary M. y David J. Unwin. 1994. *Visualization in geographical information systems*, Chichester.

Hearst, M.A. 1995. Tilebars: visualization of term distribution information in full text information access, en *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press/Addison-Wesley Publishing Co., 59–66.

Ho, Q., P. Lundblad, T. Aström, y M. Jern. 2011. A web-enabled visualization toolkit for geovisual analytics, en *Proceedings of SPIE, the International Society for Optical Engineering: SPIE: Electronic Imaging Science and Technology, Visualization and Data Analysis*, 78680R–78680R–12.

Ho, Quan Van, Patrik Lundblad, Tobias Aström, y Mikael Jern. 2012. A web-enabled visualization toolkit for geovisual analytics, *Information Visualization*, 11(1), 22–42.

Hogeweg, Marten. 2000. *Spatio-temporal visualisation and analysis*, Ph.D. thesis, University of Salford.

Huber, P.J. 1985. Projection pursuit, *The annals of Statistics*, 13(2), 435–475.

Inselberg, A. y B. Dimsdale. 1990. Parallel coordinates: A tool for visualizing multi-dimensional geometry, en *Visualization 90*, San Francisco, 361–370.

Jacobson, Ivar, Grady Booch, y James Rumbaugh. 2000. *El proceso unificado de desarrollo de software*, vol. 7, Addison Wesley Reading.

Jaramillo, Carlos Mario Zapata y Claudia Elena Durango Vanegas. 2013. Representación del conocimiento en datos espacio-temporales para sigs: un enfoque basado en esquemas preconceptuales.

Jern, Mikael, L Thygesen, y M Brezzi. 2009. A web-enabled geovisual analytics tool applied to oecd regional data, *Reviewed Proceedings in Eurographics*.

Jin, Hai y Diansheng Guo. 2009. Understanding climate change patterns with multivariate geovisualization, en *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, IEEE, 217–222.

Kandogan, E. 2000. Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions, en *Proceedings of the IEEE Information Visualization Symposium*, Citeseer, vol. 650.

Kaski, Samuel, Janne Nikkilä, y Teuvo Kohonen. 1998. Methods for interpreting a self-organized map in data analysis, en *In Proc. 6th European Symposium on Artificial Neural Networks (ESANN98). D-Facto, Brugfes*, Citeseer.

Keim, D., F. Mansmann, J. Schneidewind, J. Thomas, y H. Ziegler. 2008a. Visual analytics: Scope and challenges, *Visual Data Mining*, 76–90.

Keim, DA. 2002a. Designing pixel-oriented visualization techniques: Theory and applications, *IEEE Transactions on Visualization and Computer Graphics*, 6(1), 59–78.

Keim, DA. 2002b. Information visualization and visual data mining, *IEEE Transactions on Visualization and Computer Graphics*, 8(1), 1–8.

Keim, D.A. y H.P. Kriegel. 1994. Visdb: Database exploration using multidimensional visualization, *Computer Graphics and Applications, IEEE*, 14(5), 40–49.

Keim, D.A., F. Mansmann, J. Schneidewind, y H. Ziegler. 2006. Challenges in visual data analysis, en *Information Visualization, 2006. IV 2006. Tenth International Conference on*, IEEE, 9–16.

Keim, DA., C. Panse, y M Sips. 2005. Information visualization: Scope, techniques and opportunities for geovisualization, en J Dykes, AM MacEachren, y MJ Kraak, (Eds.) *Exploring geovisualization*, Amsterdan: Pergamon Pr, 23.

Keim, Daniel, Jorn Schneidewind, y Mike Sips. 2004. Circleview: a new approach for visualizing time-related multidimensional data sets, en ACM, (Ed.) *Proceedings of the working conference on Advanced visual interfaces*, 179–182.

Keim, Daniel A. 2000. Designing pixel-oriented visualization techniques: Theory and applications, *IEEE Transactions on Visualization and Computer Graphics*, 6.

Keim, Daniel A, Jörn Kohlhammer, Geoffrey Ellis, y Florian Mansmann. 2010. *Mastering The Information Age-Solving Problems with Visual Analytics*, Goslar, Germany: Eurographics Association.

Keim, Daniel A., Hans-Peter Kriegel, y Michael Ankerst. 1995. Recursive pattern: A technique for visualizing very large amounts of data, en *Visualization '95*, Atlanta, GA, 279–286.

Keim, Daniel A., Florian Mansmann, Jorn Schneidewind, Jim Thomas, y Hartmut Ziegler. 2008b. Visual analytics: Scope and challenges, en Simeon J. Simoff, Michael H. Bohlen, y Asturas Mazeika, (Eds.) *Visual Data Mining*, Berlin: Springer-Verlag Berlin Heidelberg, vol. 4404, 406.

Kohonen, Teuvo. 1982. Self-organized formation of topologically correct feature maps, *Biological cybernetics*, 43(1), 59–69.

Kohonen, Teuvo. 1990. The self-organizing map, *Proceedings of the IEEE*, 78(9), 1464–1480.

Korte, George. 2001. *The GIS book*, Cengage Learning.

Kraak, MJ. 2006. Visualization viewpoints: beyond geovisualization, *IEEE Computer Graphics and Applications*, 26(4), 6–9.

Levkowitz, H. 1991. Color icons-merging color and texture perception for integrated visualization of multiple parameters, en *Visualization, 1991. Visualization'91, Proceedings., IEEE Conference on*, IEEE, 164–170, 420.

Likert, Rensis. 1932. A technique for the measurement of attitudes, *Archives of psychology*.

Linstone, Harold A y Murray Turoff. 1975. *The Delphi method: Techniques and applications*, vol. 29, Addison-Wesley Reading, MA.

Long, Quan, Qingrun Zhang, Bjarni J Vilhjalmsson, Petar Forai, Ümit Seren, y Magnus Nordborg. 2013. Jawamix5: an out-of-core hdf5-based java implementation of whole-genome association studies using mixed models, *Bioinformatics*, 29(9), 1220–1222.

Luo, W., P. Yin, Q. Di, F. Hardisty, y A. M. MacEachren. 2014. A geovisual analytic approach to understanding geo-social relationships in the international trade network, *PLoS ONE*, 9(2), URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84897808830&partnerID=40&md5=982519eadc20845c38c7c51a21cd974d.

MacEachren, Alan, Xiping Dai, Frank Hardisty, Diansheng Guo, y Gene Lengerich. 2003. Exploring high-d spaces with multiform matricies and small multiples, en *International Symposium on Information Visualization*, Seattle, WA, 31–38.

MacEachren, AM y MJ Kraak. 2001. Research challenges in geovisualization, *Cartography and Geographic Information Science Special Issue on Geovisualization*, 28(1), 80.

Mazza, Riccardo. 2009. *Introduction to Information Visualization*, Springer Publishing Company.

McGrath, Robert E. 2003. Xml and scientific file formats, en *2003 Seattle Annual Meeting*.

Meservy, T. O., C. Zhang, E. T. Lee, y J. Dhaliwal. 2012. The business rules approach and its effect on software testing, en *15th International Conference on Network-Based Information Systems*, Memphis, vol. 29, 60–66.

Mitas, L, WM Brown, y H Mitasova. 1997. Role of dynamic cartography in simulations of landscape processes based on multivariate fields, *Computers & Geosciences*, 23(4), 437–446.

Mitchell, A. 1999. The esri guide to gis analysis, Environmental Systems Research Institute.

Muñiz Fernández, F., A. Carreño Torres, C. Morcillo-Suárez, y A. Navarro. 2012. Application of array-oriented scientific data formats (netcdf) to genotype data, gwaspi as an example, URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84859347982&partnerID=40&md5=d4c8938605359d9345c00547687cf71d.

Murphy, Lisa D. 1995. Geographic information systems: are they decision support systems?, en *System Sciences, 1995. Proceedings of the Twenty-Eight Hawaii International Conference on*, IEEE, vol. 4, 131–140.

NASA. 2013. Cdf user's guide, Tech. rep.

Neteler, Markus. 2010. *Open source GIS*, SAGE Publications.

Neteler, Markus, M Hamish Bowman, Martin Landa, y Markus Metz. 2012. Grass gis: A multi-purpose open source gis, *Environmental Modelling & Software*, 31, 124–130.

North, C. y B Shneiderman. 2000. Snap-together visualization: can users construct and operate coordinated visualizations?, *International Journal of Human-Computer Studies*, 53(5), 715–739.

Obeso, MaríÂa Elena Alva. 2005. *MetodologíÂa de medición y evaluación de la usabilidad en sitios web educativos*, Universidad de Oviedo.

Olaya, Victor. 2008. Sextante, a free platform for geospatial analysis, *OSGeo Journal*, 6.

Olaya, Víctor. 2011a. Introduction to sextante, Edition.

Olaya, Víctor. 2011b. Sistemas de información geográfica, *Cuadernos Internacionales de Tecnología para el Desarrollo Humano*.

Olaya, Víctor. 2011c. Sistemas de información geográfica, *Libro SIG*.

Ong, H.L. y H.Y. Lee. 1996. Software report winviz -a visual data analysis tool, *Computation & Graphics*, 20(1), 83–84.

Pfeiffer, A, I Bausch-Gall, M Otter, y Ingrid Bausch. 2012. Proposal for a standard time series file format in hdf5, en *Proceedings of 9th International Modelica Conference, Munich, Germany*.

Pickett, Ronald M. 1970. Visual analyses of texture in the detection and recognition of objects, *Picture processing and psychopictorics*, 289–308.

Pickett, Ronald M y Georges G Grinstein. 1988. Iconographic displays for visualizing multidimensional data, en *IEEE Conf. on Systems, Man and Cybernetics*, Piscataway, NJ: IEEE Press, 519.

Pirolli, Peter y Ramana Rao. 1996. Table lens as a tool for making sense of data, en ACM, (Ed.) *Proceedings of the workshop on Advanced visual interfaces*, 67–80.

Plaisant, Catherine. 2005. Information visualization and the challenge of universal usability, en *Exploring geovisualization*, Amsterdam: Elsevier, cap. Chapter 3.

Poinot, Marc. 2010. Five good reasons to use the hierarchical data format, *Computing in Science & Engineering*, 12(5), 84–90.

Ramsey, Paul. 2003. User friendly desktop internet gis (udig) for opengis spatial data infrastructures, Tech. rep., Refractions.

Rao, Ramana y Stuart K. Card. 1994. The table lens: Merging graphical and symbolic representations in an interactive focus+context visualization for tabular information, en ACM SIGCHI, (Ed.) *Conference on Human Factors in Computing Systems*.

Rew, Russ, Glenn Davis, Steve Emmerson, Harvey Davies, E Hartnett, y D Heimbigner. 2011. The netcdf users guide, data model, programming interfaces, and format for self-describing, portable data, *NetCDF Version*, 4(1).

Rhyne, Theresa Marie. 1997. Going virtual with geographic information and scientific visualization, *Computers & Geosciences*, 23(4), 489 – 491, URL http://www.sciencedirect.com/science/article/B6V7D-3T7J19D-X/2/55f21a6e4ee6ee3d60098e56ddc782e0.

Rhyne, Theresa Marie, William Ivey, Loey Knapp, Peter Kochevar, y Tom Mace. 1994. Visualization and geographic information system integration: what are the needs and requirements, if any ?, *IEEE Visualization*.

Rhyne, Theresa Marie y Alan MacEachren. 2004. Visualizing geospatial data, ACM SIGGRAPH 2004 Course 30.

Robinson, Anthony C., Jin Chen, Eugene J. Lengerich, Hans G. Meyer, y Alan M. MacEachren. 2005. Combining usability techniques to design geovisualization tools for epidemiology, *Cartography and Geographic Information Science*, 32(4), 243–255.

Rodríguez, MJ García, A Urrutia Zambrana, y MA Bernabé Poveda. 2009. Diseño de herramientas de análisis espacio-temporales para el estudio de bases de datos históricas, en *VI Jornadas Técnicas de la IDE de España*.

Rogowitz, Bernice E. y Lloyd A. Treinish. 1993. An architecture for perceptual rule-based visualization, en *IEEE Visualization 93 Proceedings*, 236 – 243.

Sacha, Dominik, Andreas Stoffel, Florian Stoffel, Bum Kwon, Geoffrey Ellis, y Daniel Keim. 2014. Knowledge generation model for visual analytics, *Visualization and Computer Graphics, IEEE Transactions on*, 20(12), 1604 – 1613.

Sagl, G. y E. Delmelle. 2014. Mapping collective human activity in an urban environment based on mobile phone data, *Cartography and Geographic Information Science*, URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84900463220&partnerID=40&md5=8050a766ef6c656de8fffb60978f4ad1.

Shekhar, Shashi y Hui Xiong. 2008. *Encyclopedia of GIS*, Springer.

Shneiderman, B. 1996. The eyes have it: A task by data type taxonomy for information visualizations, en *IEEE Symposium on Visual Languages*, IEEE, 336–343.

Spence, B., L. Tweedie, H. Dawkes, y H. Su. 1995. Visualisation for functional design, IEEE, 4–10.

Star, Jeffrey y John Estes. 1990. *Geographic information systems*, prentice-Hall Englewood Cliffs.

Steiniger, Stefan y Andrew JS Hunter. 2012. Free and open source gis software for building a spatial data infrastructure, en *Geospatial free and open source software in the 21st century*, Springer, 247–261.

Strahler, A.H. y A.N. Strahler. 1992. *Modern physical geography.*, New York, USA, 4th edn.

Streit, Marc, Rupert C Ecker, Katja Osterreicher, Georg E Steiner, Horst Bischof, Christine Bangert, Tamara Kopp, y Radu Rogojanu. 2006. 3d parallel coordinate systems—a new data visualization method in the context of microscopy-based multicolor tissue cytometry, *Cytometry Part A*, 69(7), 601–611.

Sundmaeker, Harald, Patrick Guillemin, Peter Friess, y Sylvie Woelfflé. 2010. *Vision and challenges for realising the Internet of Things*, EUR-OP.

Symanzik, Jürgen, Dianne Cook, Nicholas Lewin-Koh, James J Majure, y Inna Megretskaia. 2000. Linking arcview and xgobi: Insight behind the front end, *Journal of Computational and Graphical Statistics*, 9(3), 470–490.

Takatsuka, Masahiro y Mark Gahegan. 2002. Geovista studio: a codeless visual programming environment for geoscientific data analysis and visualization, *Computers & Geosciences*, 28(10), 1131–1144.

Theisel, H. 2000. Scientific visualization, URL http://www.aimatshape.net/s05_sciviz.zip.

Theus, Martin. 2005. Exploring geovisualization, chapter 6. statistical data exploration and geographical information visualization, en *Exploring geovisualization*, Amsterdam: Elsevier, cap. Chapter 6.

Tominski, Christian, James Abello, y Heidrun Schumann. 2004. Axes-based visualizations with radial layouts, en *Proceedings of the 2004 ACM symposium on Applied computing*, ACM, 1242–1247.

Tukey, J.W. 1977. *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.

Turton, Ian. 2008. Geo tools, en *Open source approaches in spatial data handling*, Springer, 153–169.

Ullman, Richard y Michael Denning. 2012. Hdf5 for npp sensor and environmental data records, en *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*, IEEE, 1100–1103.

UNIDATA. 2012. URL http://www.unidata.ucar.edu.

Valle-Lima, A. D. 2012. *La investigación pedagógica. Otra mirada*, Pueblo y Educación.

Vanegas, Claudia Elena Durango. 2013. Caracterización de datos espacio-temporales en sistemas de información geográfica.

Vázquez-Rodríguez, R., C. Pérez-Risquet, I. Y. Gonzalez-Herrera, A. Fajardo-Moya, y J. C. Torres-Cantero. 2010a. A new visual data mining tool for gvsig gis, en *KDIR 2010 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, 428–431, URL http://www.scopus.com/inward/record.url?eid=2-s2.0-78651412580&partnerID=40&md5=41d740c6f504e4f4ee17f53b9cfbfd20.

Vázquez-Rodríguez, Romel, Inti González-Herrera, y Carlos Pérez Risquet. 2009a. Estudio de la factibilidad de la aplicación de técnicas de visualización de datos multiparamétricos para el análisis visual de datos meteorológicos, en *CompuMat 2009*, La Habana.

Vázquez-Rodríguez, Romel, Carlos Pérez-Risquet, Inti González-Herrera, Alexis Fajardo-Moya, Reinier Oves-Garcia, y Juan Carlos Torres-Cantero. 2011a. Incorporación de herramientas de visualización para el análisis visual de datos en sig de escritorio, en *Informática 2011*, La Habana, Cuba.

Vázquez-Rodríguez, Romel, Carlos Pérez-Risquet, Inti González-Herrera, Juan Carlos Torres-Cantero, y Alexis Fajardo-Moya. 2009b. Implementación de técnicas de visualización científica para el análisis visual de datos meteorológicos, en *RecPat 2009*, Santiago de Cuba, Cuba.

Vázquez-Rodríguez, Romel, Carlos Pérez-Risquet, Inti González-Herrera, Juan Carlos Torres-Cantero, y Alexis Fajardo-Moya. 2010b. Uso de técnicas de visualización científica en sig para el análisis visual de datos meteorológicos, *Serie Científica Universidad de las Ciencias Informáticas*, 3(6).

Vázquez-Rodríguez, Romel, Carlos Pérez-Risquet, Inti González-Herrera, Juan Carlos Torres-Cantero, Alexis Fajardo-Moya, y Reinier Oves-Garcia. 2011b. Visualización espacio-temporal mediante técnicas de visualización de datos multiparamétricos en sig, en *Geociencias 2011*, La Habana, Cuba.

Vázquez-Rodríguez, Romel, Carlos Pérez-Risquet, Inti González-Herrera, Juan Carlos Torres-Cantero, y Alexis Fajardo Moya. 2010c. Implementation of scientific visualization techniques for meteorological visual data in gis, en *CFWMLKD2010*, Santa Clara, Cuba.

Vázquez-Rodríguez, Romel, Carlos Pérez-Risquet, Inti Y. Gonzalez-Herrera, Alexis Fajardo-Moya, y Juan C. Torres-Cantero. 2010d. A new visual data mining tool for gvsig gis, en INSTICC, (Ed.) *International Conference on Knowledge Discovery and Information Retrieval*, Valencia, Spain: INSTICC.

Vázquez-Rodríguez, Romel, Carlos Pérez-Risquet, y Reinier Oves-Garcia. 2009c. Uso de sistemas de información geográfica basados en software libre para la visualización de datos meteorológicos, en *Informática 2009*, Habana, Cuba.

Vázquez-Rodríguez, Romel, Carlos Pérez-Risquet, y Juan Carlos Torres. 2015. Exploratory data analysis through the integration of visualization techniques in geographical information systems, *Revista Técnica de la Facultad de Ingeniería de la Universidad del Zulia*, 38(1), 73–82.

Vázquez-Rodríguez, Romel, Carlos Pérez-Risquet, y Juan Carlos Torres-Cantero. 2009d. Visualización de datos meteorológicos mediante técnicas de visualización científica en sig, en *Tendencias en Soft-computing*, Santa Clara: Seminario Internacional de Doctorado en Soft Computing. Santa Clara, Cuba. 30, Marzo 2009. ISBN: 959250525-4.

Vázquez-Rodríguez, Romel, Carlos Pérez-Risquet, y Juan Carlos Torres-Cantero. 2013a. Extensión del módulo de visualización de datos espacio-temporales de gvsig, en *Informática 2013*, La Habana, Cuba.

Vázquez-Rodríguez, Romel, Carlos Pérez-Risquet, y Juan Carlos Torres-Cantero. 2013b. A novel visual data mining module for the geographical information system gvsig, *Anuário do Instituto de Geociências*, 36(1), 98–111.

Vázquez-Rodríguez, Romel, Carlos Pérez-Risquet, Juan Carlos Torres-Cantero, y Alexis Gallardo-Segura. 2013c. Integración de formatos de datos científicos en sistemas de información geográfica, en *Geociencias 2013*, La Habana, Cuba.

Vázquez-Rodríguez, Romel, Carlos Pérez-Risquet, Juan Carlos Torres-Cantero, y Rainer Martínez-Fraga. 2013d. Nuevas técnicas de visualización para el módulo de visualización de datos espacio-temporales de gvsig, en *Geociencias 2013*, La Habana, Cuba.

Wang, Y., Y. Su, y G. Agrawal. 2013. Supporting a light-weight data management layer over hdf5, en *Proceedings - 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing, CCGrid 2013*, 335–342, URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84881277221&partnerID=40&md5=fbc950392d86fb0b0b4b8fd43cdee5f0.

Ward, Matthew O. 2002. *A Taxonomy of Glyp Placement Strategies for Multidimensional Data Visualization*, Ph.D. thesis, Worcester Polytechnic Institute.

Weber, M., M. Alexa, y W. Müller. 2001. Visualizing time-series on spirals, en *Information Visualization*, Citeseer, 7.

Wijk, J.J. van y R.D. van Liere. 1993. Hyperslice, en *IEEE Visualization '93*, Los Alamitos: IEEE Computer Society Press, 119–125.

Wildman, P. 2005. Stat 2005: An internet course in statistics, URL http://wind.cc.whecn.edu/~pwildman/statnew/information.htm.

Wong, Pak Chung y R Daniel Bergeron. 1994. 30 years of multidimensional multivariate visualization, en *Scientific Visualization*, 3–33.

Wright, William. 1995. Research report: Information animation applications in the capital markets, en *Information Visualization, 1995. Proceedings.*, IEEE, 19–25.

Xie, Z, S Huang, MO Ward, y EA Rundensteiner. 2006. Exploratory visualization of multivariate data with variable quality, en *IEEE Symposium On*, IEEE, 183–190.

Yang, Wenli y Liping Di. 2004. An accurate and automated approach to georectification of hdf-eos swath data, *Photogrammetric engineering and remote sensing*, 70(4), 397–404.

Yeh, Pen-Shu, Wei Xia-Serafino, Lowell Miles, Ben Kobler, y Daniel Menasce. 2002. Implementation of ccsds lossless data compression in hdf, en *Earth Science Technology Conference*.

Yi, Ji Soo, Youn ah Kang, John T Stasko, y Julie A Jacko. 2007. Toward a deeper understanding of the role of interaction in information visualization, *Visualization and Computer Graphics, IEEE Transactions on*, 13(6), 1224–1231.

Yuan, M. 1996. Temporal gis and spatio-temporal modeling, en *Proceedings of Third International Conference Workshop on Integrating GIS and Environment Modeling*, Santa Fe, NM.

Zhang, Z., X. Tong, K. T. McDonnell, A. Zelenyuk, D. Imre, y K. Mueller. 2013. An interactive visual analytics framework for multi-field data in a geo-spatial context, *Tsinghua Science and Technology*, 18(2), 111–124, URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84897584756&partnerID=40&md5=71138433c9a4965007fdb6a4b20bff12.

Zhao, S., T. Yu, Q. Meng, Q. Zhou, F. Wang, L. Wang, y Y. Hu. 2010. Gdal-based extend arcgis engine's support for hdf file format, en *Geoinformatics, 2010 18th International Conference on*, IEEE, 1–3.

Anexo 1. Manual de usuario extScientificVisualization 1.0

Para utilizar la extensión extScientificVisualization 1.0 es necesario copiar el módulo compilado para el directorio de extensiones de gvSIG. Se recomienda utilizar la versión 1.9 de gvSIG, ya sea la versión portable o instalable.

El análisis visual de datos multiparamétricos utilizando las técnicas de visualización con que cuenta la extensión implementada, requiere un conocimiento previo de cómo aplicarlas de forma adecuada para lograr una correcta comprensión de los resultados que se obtengan.

Esta extensión realizada a gvSIG permite la exploración y análisis visual de datos multiparamétricos mediante las técnicas: coordenadas paralelas, gráfico de Andrews, técnicas basadas en iconos, segmentos de círculo y patrones recursivos. Permite visualizar un único conjunto de datos donde los parámetros de la visualización dependen solamente de los valores de las variables del propio conjunto, así como visualizar coordinadamente varios conjuntos de datos donde los atributos de la visualización representan el comportamiento global de las variables para la totalidad de los datos.

Visualización de un conjunto de datos

El especialista puede realizar la visualización de un único conjunto de datos que estén previamente configurados en los formatos de archivos de datos *arff* y *dbf*. Para realizar este análisis se puede acceder a una opción del menú *Visualization/Open and visualize data* o a un botón de la barra de herramientas de gvSIG. En la figura A.1 se pueden observar resaltadas en rojo las respectivas opciones.

Al seleccionar algunas de estas opciones se muestra una vista para la selección del archivo de datos, como la en la figura A.2. Resaltado en rojo con el número uno está el botón que muestra un diálogo de selección de ficheros que permite abrir solo los formatos de archivos permitidos. Luego de ser seleccionado el archivo de datos se muestra su dirección en el campo de texto resaltado con el número dos y se habilita el botón *Next* resaltado con el número tres.

Al pulsar el botón *Next* se muestra la vista para la selección de las técnicas de visualización,

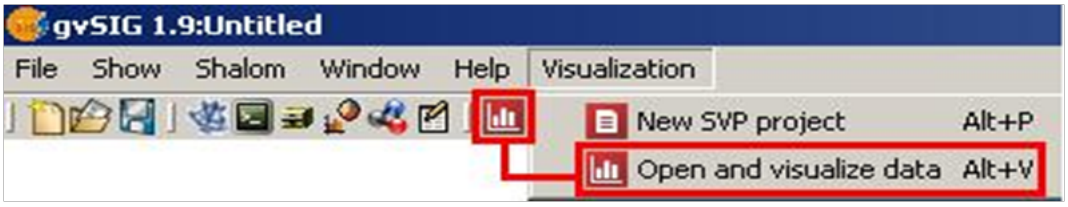

Figura A.1 Opciones para realizar la visualización de un conjunto de datos.

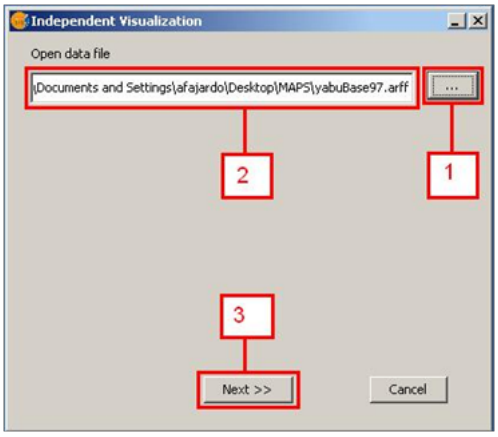

Figura A.2 Vista para la selección del fichero de datos.

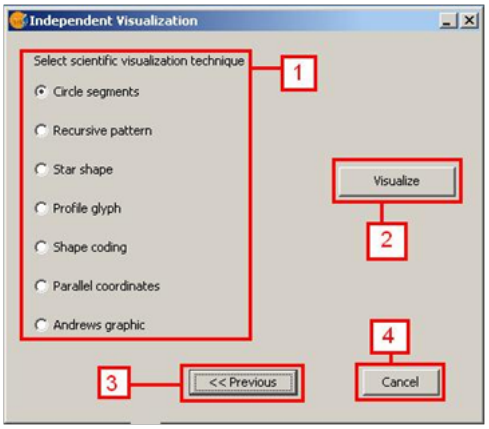

Figura A.3 Vista para la selección y visualización de las técnicas.

obsérvese la figura A.3.

La región identificada con el número uno muestra las técnicas con las que puede ser visualizado el conjunto de datos seleccionado. Una vez seleccionada alguna de las opciones se procede a pulsar el botón *Visualize*, número dos, que muestra la representación visual de los datos con la técnica de visualización escogida. De esta manera se puede visualizar el mismo conjunto de datos con todas las técnicas que desee el usuario. En caso de que se desee visualizar otro conjunto de datos distinto del seleccionado se pulsa al botón *Previous*, número tres, retornando a la vista de selección del archivo de datos de la figura A.2. El botón *Cancel*, número cuatro, cierra la ventana.

Coordenadas paralelas

En la figura A.4 se puede observar un ejemplo de la aplicación de la técnica coordenadas paralelas, que será utilizada para explicar las opciones de configuración que brinda la técnica.

La implementación de esta técnica permite intercambiar los ejes que representan los atributos, lo que posibilita que puedan ser reorganizados de acuerdo con las necesidades del usuario. El procedimiento consiste en posicionar el puntero del "ratón" encima de un eje coordenado, hacer un *clic* izquierdo, trasladarse a otro eje y seleccionarlo mediante un *clic*; luego de realizar esta operación las barras se intercambian y con ello cambia el orden de los atributos en la imagen. A continuación se exponen las opciones enumeradas en la figura:

1. *High quality*: Habilita trazos mejor definidos de las líneas aumentado la calidad de la imagen.

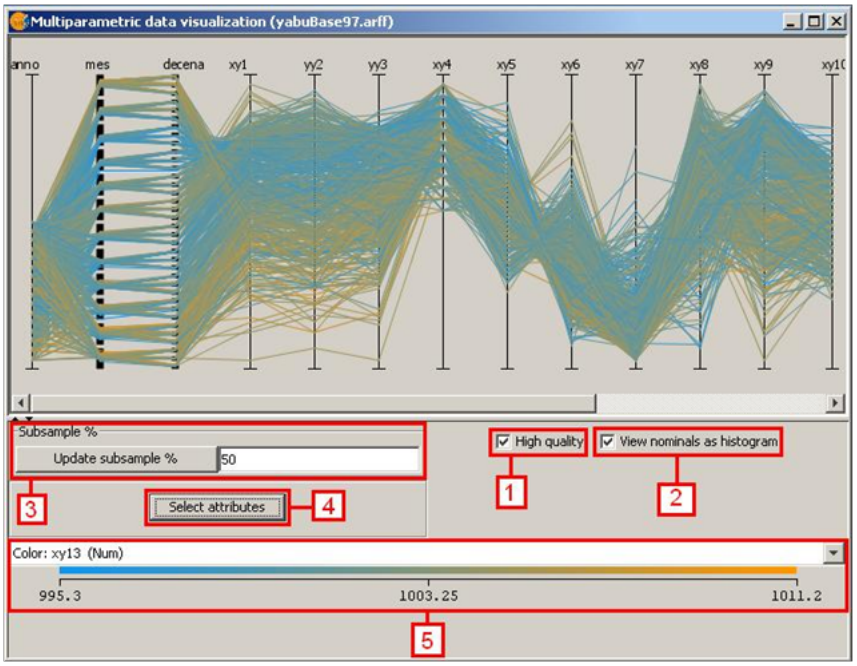

Figura A.4 Visualización de un conjunto de datos utilizando la técnica coordenadas paralelas.

Figura A.5 Diálogo de selección de atributos.

2. *View nominals as histogram*: Las variables cuyo dominio de definición es un conjunto finito de valores pueden provocar superposiciones en la imagen de la visualización. Seleccionando esta opción el eje se divide en tantas barras como valores tenga la variable que representa, haciendo corresponder cada barra a un valor específico del atributo.
3. *Update subsample %*: Esta opción permite obtener una muestra del conjunto de datos dado el por ciento insertado en el campo de texto.
4. *Select attributes*: Al elegir esta opción se muestra el diálogo de la figura A.5 que permite la selección de los atributos que se deseen visualizar.
5. Selección del color de las observaciones: Esta opción permite asignarle el color a las observaciones con respecto a los valores de un atributo. Para ello se debe seleccionar un atributo y automáticamente se muestra su escala de colores, redibujándose las observaciones de acuerdo con el valor que posee la dimensión en dicha observación.

Gráfico de Andrews

La técnica gráfico de Andrews presenta, al igual que la anterior, las opciones de *Update Subsample %* y *Select Attributes*. Estas opciones pueden resultar en extremo útiles si el número

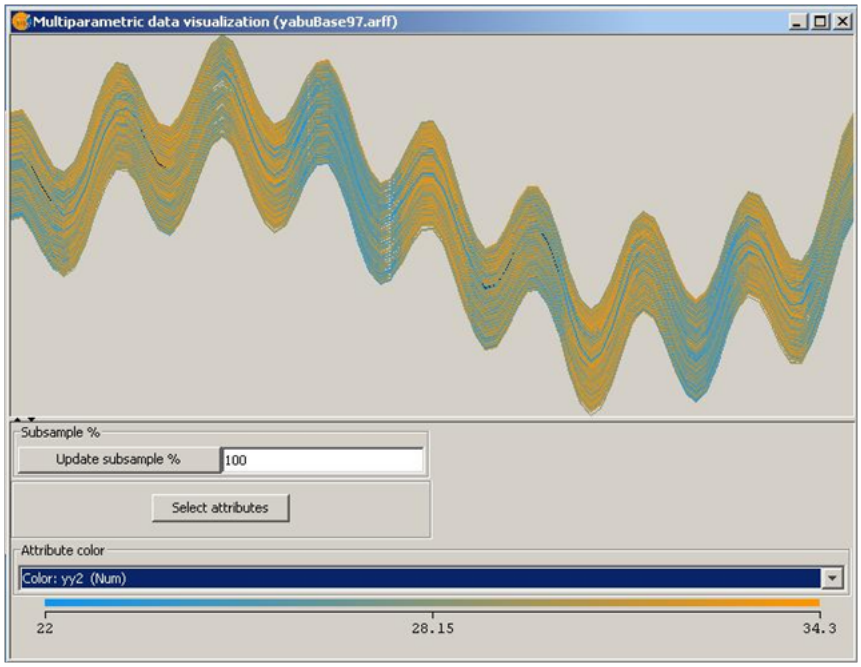

Figura A.6 Visualización de un conjunto de datos empleando la técnica gráfico de Andrews.

de observaciones es grande o si existen demasiados atributos que no resultan de interés, ya que disminuye el desorden en la imagen.

La técnica permite definir el color de cada observación. El esquema es similar al utilizado en la técnica coordenadas paralelas anteriormente descrita, que se basa en mantener un atributo para asignar el color de las observaciones mediante una escala de colores. La utilización de esta técnica se puede apreciar en la figura A.6.

Técnicas basadas en iconos

La técnica basada en iconos implementada sigue el modelo de la técnica campo de estrellas. Se trata de visualizar los iconos en un espacio siguiendo un método de localización. En la actual implementación la localización de los iconos se realiza atendiendo al orden natural de los datos. La figura A.7 muestra la utilización de la técnica con iconos en forma de estrella.

De manera análoga a las técnicas anteriores, el usuario puede visualizar muestras de observaciones dado un por ciento y seleccionar atributos. Permite aumentar el tamaño de los iconos mediante la barra de desplazamiento resaltada en color rojo. En el caso de los iconos en forma de estrella (obsérvese la figura A.7) e iconos en forma de barras (obsérvese la figura A.8) los colores de los iconos, que representan las observaciones, pueden ser visualizados con un color de acuerdo con los valores de un atributo, de manera similar a la explicación brindada para coordenadas paralelas. En el caso de shapecoding (obsérvese la figura A.9), esta operación no es posible puesto que los colores de cada celda dependen de los valores del atributo que representa; la opción *Attribute color* puede ser utilizada a manera de leyenda, se puede apreciar el rango de valores y la codificación de esos valores en una escala de colores para cada atributo.

Segmentos de círculo

La utilización de esta técnica puede apreciarse en la figura A.10. Esta visualización también brinda las opciones de seleccionar los atributos de interés para el usuario y la obtención de muestras del total de observaciones dado un por ciento. Aunque esta técnica está especialmente diseñada para mostrar un gran volumen de datos, en caso de que se tenga un conjunto extremadamente grande de los mismos, sí se debe considerar la reducción de las observaciones. La cantidad de atributos suele influir mucho más en la expresividad de la representación, pues

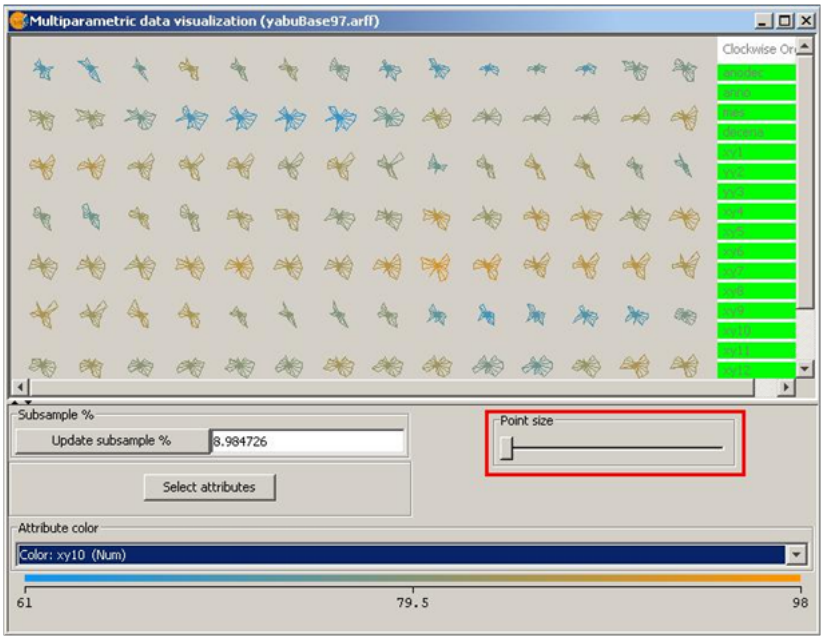

Figura A.7 Icono en forma de estrella.

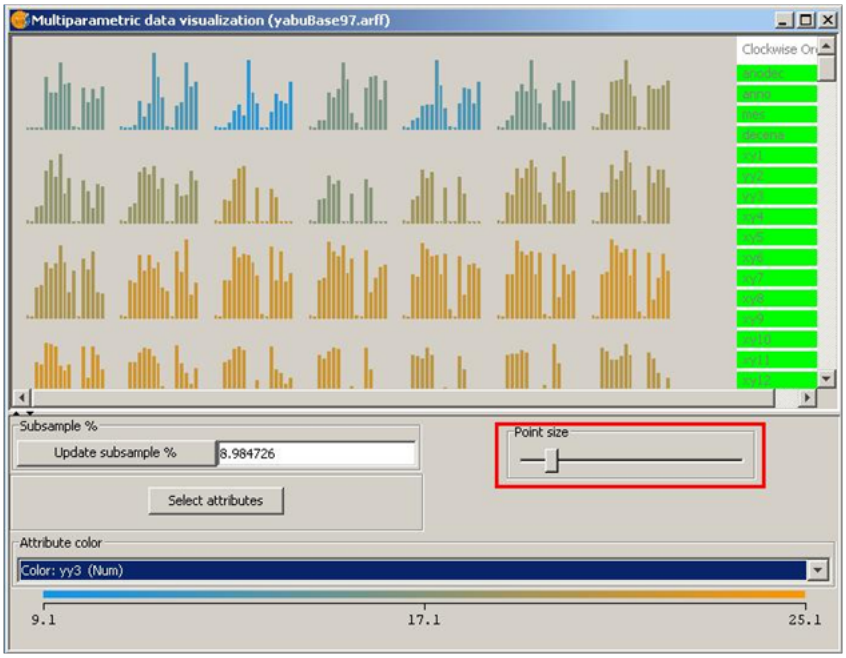

Figura A.8 Visualización de un conjunto de datos empleando iconos en forma de barras.

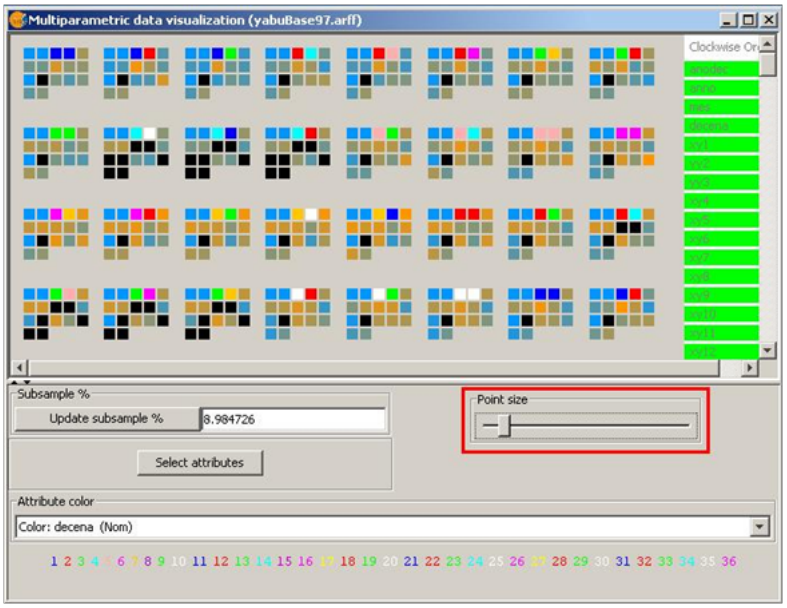

Figura A.9 Visualización de un conjunto de datos empleando *shapecoding*.

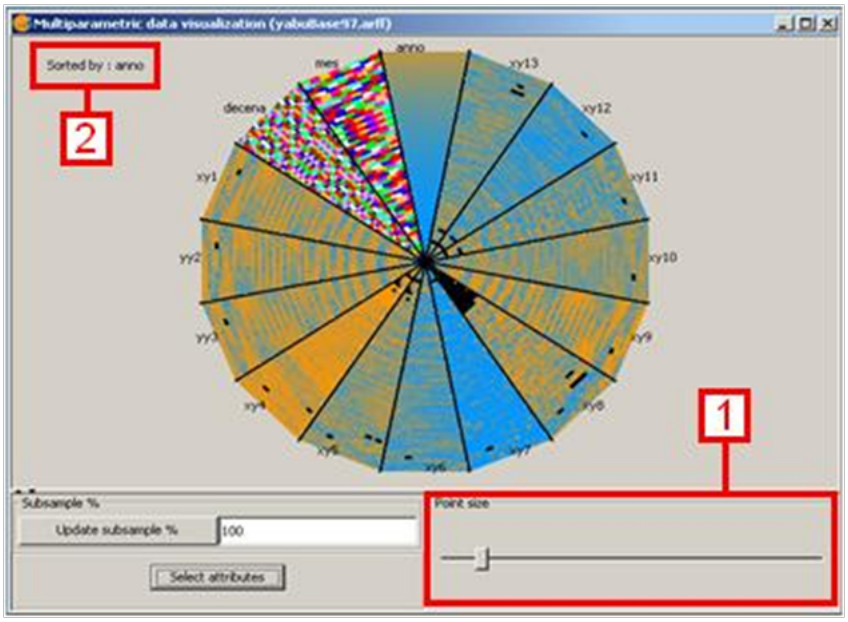

Figura A.10 Técnica segmentos de círculo.

un número elevado de estos disminuye en gran medida el espacio de representación de cada atributo, en cuyo caso lo mejor es realizar una selección de los mismos.

Otra posible interacción que se brinda es la elección del atributo mediante el cual se establece el orden de las observaciones. Por defecto las mismas aparecen ordenadas por el primer atributo que aparece en los datos, pero el usuario puede elegir cualquier otro con solo seleccionar el atributo deseado haciendo *clic* en la etiqueta correspondiente sobre la imagen, al realizar esta operación se visualiza el atributo seleccionado en la región resaltada en color rojo con numeración dos. Otro de los parámetros de interacción que se brinda es la barra de desplazamiento situada en la esquina inferior derecha con numeración uno, la cual permite al usuario controlar el tamaño de los puntos que representan los valores de cada variable en una observación.

Patrones recursivos

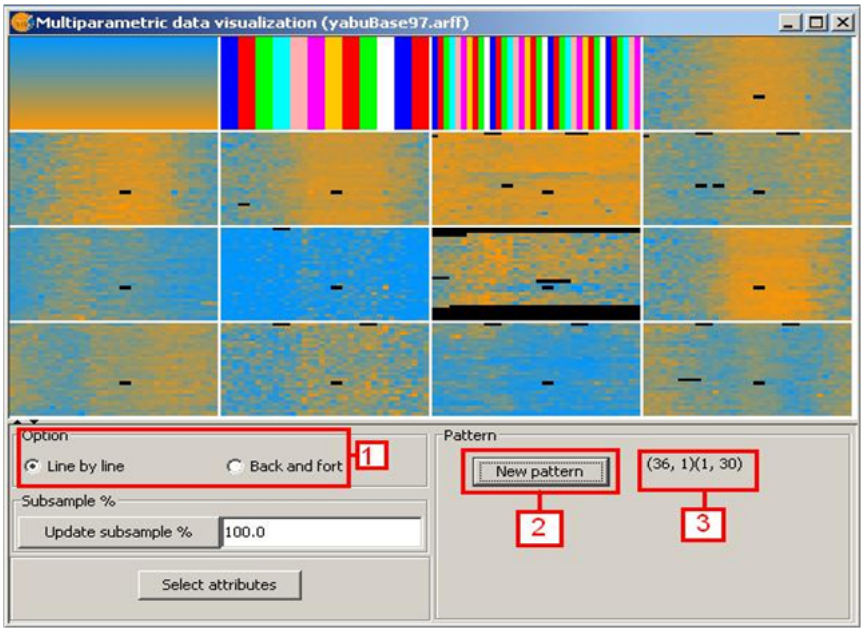

Figura A.11 Visualización de un conjunto de datos mediante la técnica patrones recursivos.

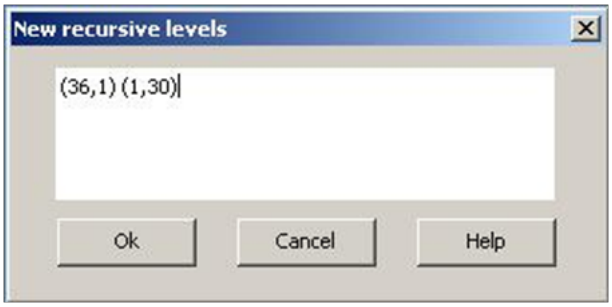

Figura A.12 Diálogo para la edición de un nuevo patrón.

La utilización de la técnica patrones recursivos puede ser apreciada en la figura A.11. Esta técnica permite realizar las operaciones *Update Subsample %* y *Select Attributes* explicadas anteriormente.

La opción identificada con el número uno permite personalizar la forma de posicionar los píxeles en la región particular de cada atributo. Seleccionando *Line by line* los píxeles son posicionados línea a línea, de izquierda a derecha para cada nivel de recursividad; si se selecciona *Back and fort* los píxeles son posicionados de izquierda a derecha en una línea y en la siguiente de derecha a izquierda, repitiendo el procedimiento para cada nivel de recursividad. La opción dos es el botón *New pattern*, que al ser pulsado muestra el diálogo de la figura A.12 para la edición de los niveles de recursividad que conforman un nuevo patrón.

La forma de editar un nuevo patrón es definiendo los niveles de recursividad del mismo. Los niveles de recursividad se representan por pares de números enteros entre paréntesis y separados por coma; los niveles de recursividad se separan por un espacio entre los pares de valores.

Según esta estructura a esta estructura suponga un ejemplo en que se tienen 120 observaciones de distintas variables obtenidas mensualmente durante diez años; en este caso el orden de los datos lo proporcionaría la variable Mes o el orden natural de los datos si las observaciones poseen un orden cronológico. Se puede obtener el comportamiento anual, durante los diez años para todas las variables insertando el patrón $(12, 1)\ (1, 10)$, donde el nivel de recursividad $(12, 1)$ corresponde a posicionar 12 mediciones de las variables en una línea y el nivel $(1, 10)$ repetir el

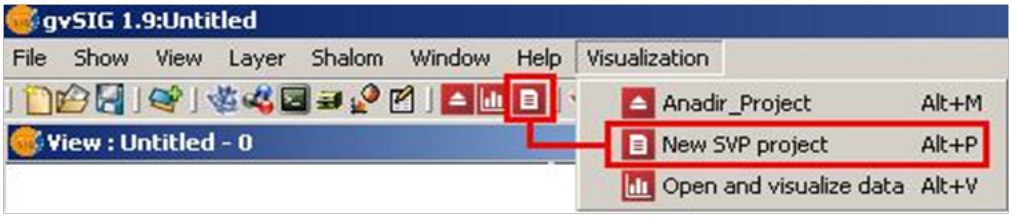

Figura A.13 Opciones para la configuración de un nuevo proyecto de visualización coordinada.

procedimiento anterior 10 veces de arriba hacia abajo, por lo tanto se estarían representando las $12 * 10 = 120$ observaciones de las variables. Es responsabilidad del usuario decidir qué patrón utilizar, para ello debe considerar la cantidad de observaciones que se deseen visualizar, el orden natural de los datos, aunque se puede utilizar una variable que represente un orden temporal y todo el conocimiento que pueda aportar el usuario sobre los datos.

El patrón que está siendo utilizado por la visualización se muestra en el campo de texto señalado con el número tres en la figura A.11.

Visualización coordinada de varios conjuntos de datos

La configuración de un proyecto de visualización científica puede ser realizada accediendo a las opciones que se muestran en la figura A.13.

Esta operación se puede realizar a través de un botón en la barra de herramientas o accediendo a la opción de menú *Visualization/New SVP project*, como se muestra en la figura A.13. Al pulsar una de las opciones antes descritas se muestra un asistente para la configuración del proyecto. El asistente permite realizar la descripción del proyecto que se desea crear, para conocimiento del usuario en caso de que desee utilizar en otro momento un proyecto creado con anterioridad. Una vez realizada la descripción del proyecto (opcional) el usuario puede seleccionar el directorio donde guardar el proyecto así como el nombre del archivo con extensión *svp* donde se guardará la configuración del mismo.

A continuación se selecciona el mapa que servirá como base o fondo a la visualización y se procede a seleccionar el mapa que será utilizado para las localizaciones, que puede ser el mapa base o un mapa con las localizaciones exactas del origen de los datos que tenga total correspondencia con el mapa base. Realizada la anterior operación, el usuario define el campo de la tabla asociada al mapa de las localizaciones por el cual referenciará los archivos con los datos; a continuación el usuario puede seleccionar la localización en el mapa y el archivo de datos que desea asociar a la localización, realizando tantas inserciones como localizaciones tenga el mapa.

Cada archivo de datos asociado a una localización será visualizado en el centroide de la geometría correspondiente. El asistente controla, además, que los ficheros con los datos no sean utilizados en más de una localización. En el **Anexo 2** se muestran ejemplos de ficheros de configuración *svp*.

Al seleccionar una de las opciones para la configuración de un proyecto de visualización coordinada se muestra un asistente cuya ventana inicial se puede observar en la figura A.14.

La descripción de la vista es como sigue a continuación, según la numeración en la figura:

1. Campo de texto que permite insertar una descripción del proyecto. Esta operación es opcional.
2. Botón *Next*: carga la siguiente vista que se muestra en la figura A.15.
3. Botón *Cancel*: cancela la creación del proyecto.

Esta vista permite salvar el fichero de configuración del proyecto con un nombre y en el directorio especificado por el usuario. Al pulsar el botón dos se muestra un diálogo para salvar

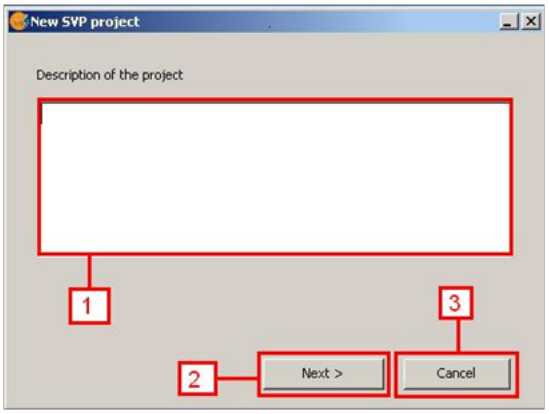

Figura A.14 Vista para añadir la descripción del proyecto.

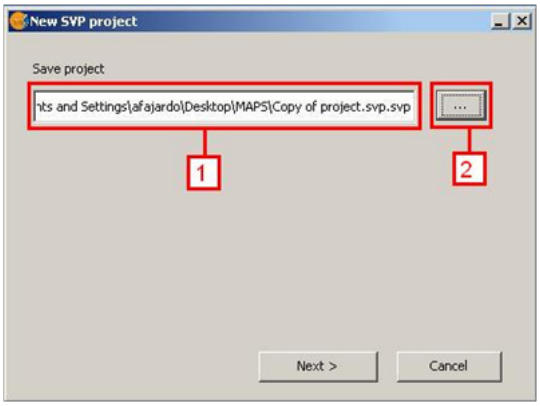

Figura A.15 Selección del directorio del proyecto de visualización coordinada.

el fichero, una vez aceptada la selección se muestra en el campo de texto uno la dirección donde se salvará el fichero. Al pulsar el botón *Next* se muestra la vista de la figura A.16.

Esta vista permite la selección, mediante el botón dos, del mapa base para la visualización de las técnicas. La dirección y nombre del archivo se muestran en uno. Una vez realizada la selección se pulsa el botón *Next* y se muestra la vista de la figura A.17 para la selección del mapa que será utilizado para las localizaciones.

Las opciones enumeradas se explican a continuación:

1. Selección del mapa para las localizaciones: esta opción le permite al usuario definir el ma-

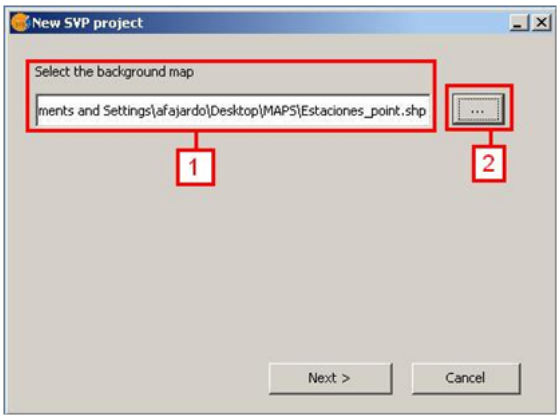

Figura A.16 Selección del mapa base.

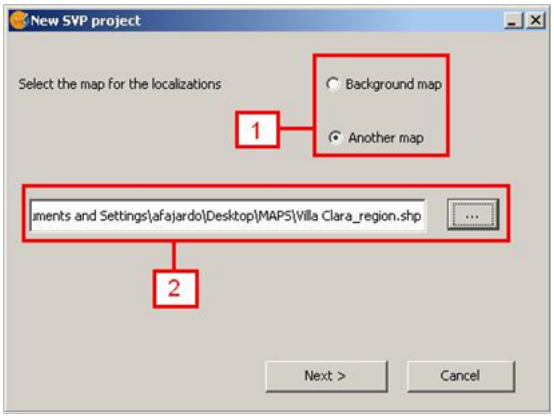

Figura A.17 Selección del mapa para las localizaciones.

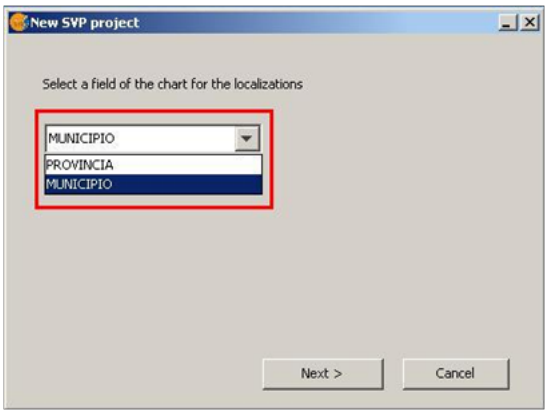

Figura A.18 Selección del campo para las localizaciones.

pa que será utilizado para realizar las localizaciones de los archivos de datos. Por defecto la opción activada es *Background map*, que significa que la localización será realizada sobre el mapa base seleccionado en la vista anterior. Por el contrario, si elige la opción *Another map*, se habilitan el campo de texto y el botón con numeración dos.

2. Selección de otro mapa para las localizaciones: Al ser seleccionada la opción *Another map* el usuario tiene la posibilidad de seleccionar un mapa vectorial de puntos que contenga las localizaciones. Este nuevo mapa debe tener una correspondencia con el mapa base; principalmente se utiliza esta opción cuando se dispone de un mapa con las localizaciones exactas en la región del origen de los datos que se deseen visualizar.

Al pulsar el botón *Next* se muestra la vista de selección del atributo del mapa de localizaciones que el usuario utilizará para localizar los archivos de datos. Obsérvese la figura A.18.

Todo archivo en formato ESRI **shape** tiene asociado un archivo en formato de tabla que contiene información asociada a los objetos geométricos, como pueden ser los nombres de las regiones, coordenadas de las localizaciones, etc. Esta tabla tiene tantas filas como geometrías tenga definido el mapa. Las geometrías a su vez pueden ser líneas, puntos o polígonos. Es posible asociarle un archivo de datos multiparamétricos a una geometría que representa una localización. La correcta selección del atributo de la tabla que servirá como guía para localizar los datos es fundamental, pues permite conocer con mayor certeza a qué región o punto del mapa se está asociando el archivo con los datos multiparamétricos. Una vez seleccionado el atributo se pulsa el botón *Next* y se muestra la vista de la figura A.19.

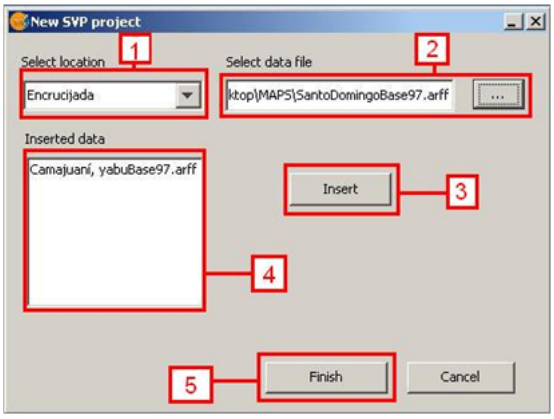

Figura A.19 Inserción de los archivos de datos en las localizaciones.

Esta vista permite asociar un archivo de datos multiparamétricos a una localización del mapa. A continuación se brinda la descripción de las componentes que conforman la vista:

1. Selección de la localización a la que se va a asociar un archivo de datos. Se muestran los valores según el campo seleccionado en el paso anterior.
2. Opción que muestra un diálogo para la selección del fichero de datos asociado a la localización seleccionada en uno.
3. Botón *Insert*: inserta la localización y el nombre del fichero de datos en el archivo de configuración del proyecto.
4. Campo de texto que muestra las inserciones realizadas.
5. Botón *Finish*: termina la configuración del proyecto de visualización coordinada y cierra la vista.

Utilización de un proyecto de visualización coordinada

Cuando se dispone de un proyecto previamente creado el primer paso para realizar la visualización lo constituye crear una vista de visualización.

La figura A.20 muestra los pasos para la creación de una vista de visualización coordinada, los cuales son: seleccionar el tipo de documento de visualización, identificado en la figura con el número uno, crear un nuevo documento de visualización oprimiendo el botón *New* (número dos), seleccionar el documento creado (número tres) y abrirlo oprimiendo el botón *Open* con numeración cuatro. Realizados estos pasos se muestra una vista como la resaltada con el número cinco.

Al estar activa una vista de visualización se habilita una opción de menú *Visulization/Load SVP project* y un botón en la barra de herramientas, como lo muestra la figura A.21.

Cuando se selecciona una de las opciones anteriores, se muestra un diálogo para la selección del fichero *svp*, que constituye el proyecto de visualización coordinada con el que se desea trabajar. Una vez seleccionado el archivo, se carga el proyecto en la vista de visualización activa y automáticamente se cargan los mapas que estén en el archivo de configuración.

La figura A.22 muestra los diferentes componentes que conforman una vista de visualización. Resaltado en color rojo se muestra la opción de selección de la técnica de visualización a utilizar, esta opción consta de dos componentes: un *combobox* para la selección de las técnicas de visualización y el botón *Visualize* para realizar la visualización de la técnica seleccionada. Al oprimir el botón *Visualize* se muestra en el área resaltada con color azul el panel de configuración de la técnica con que se está realizando la visualización de los datos. La región resaltada en color verde constituye el área de visualización de los mapas y de las técnicas que pueden ser

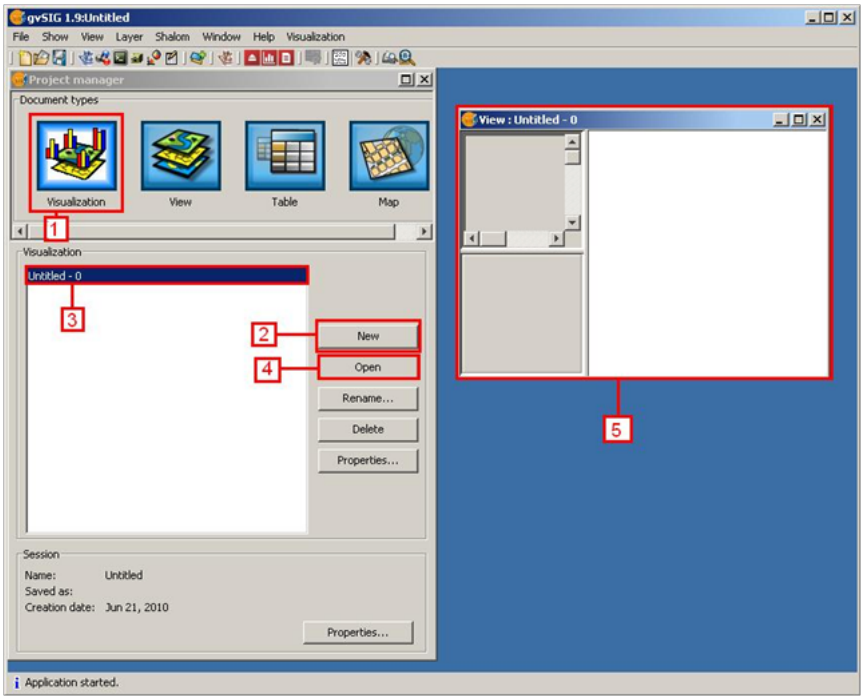

Figura A.20 Creación de una vista de visualización coordinada.

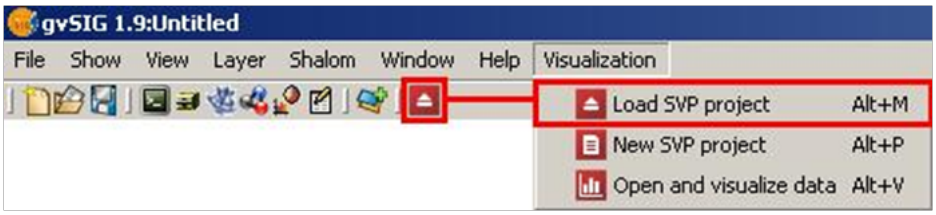

Figura A.21 Opciones para la adición de un proyecto de visualización coordinada a una vista de visualización.

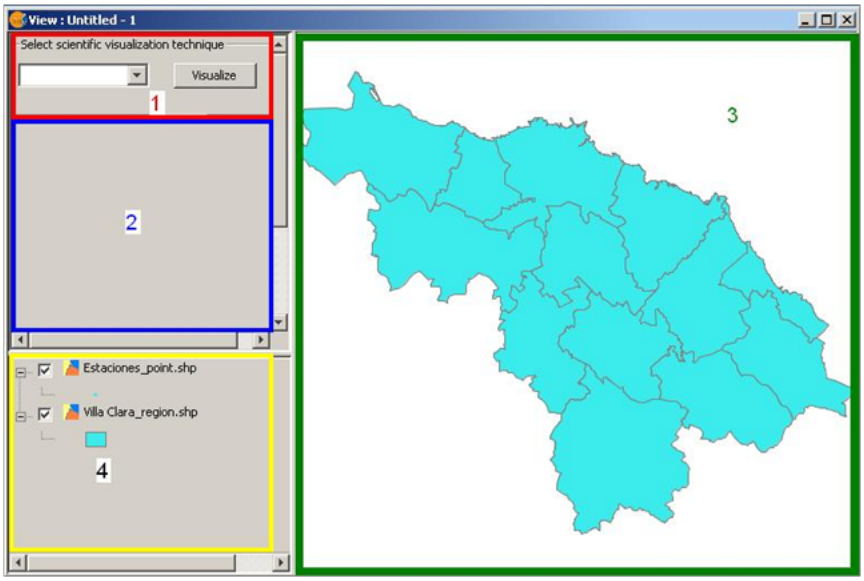

Figura A.22 Vista de visualización.

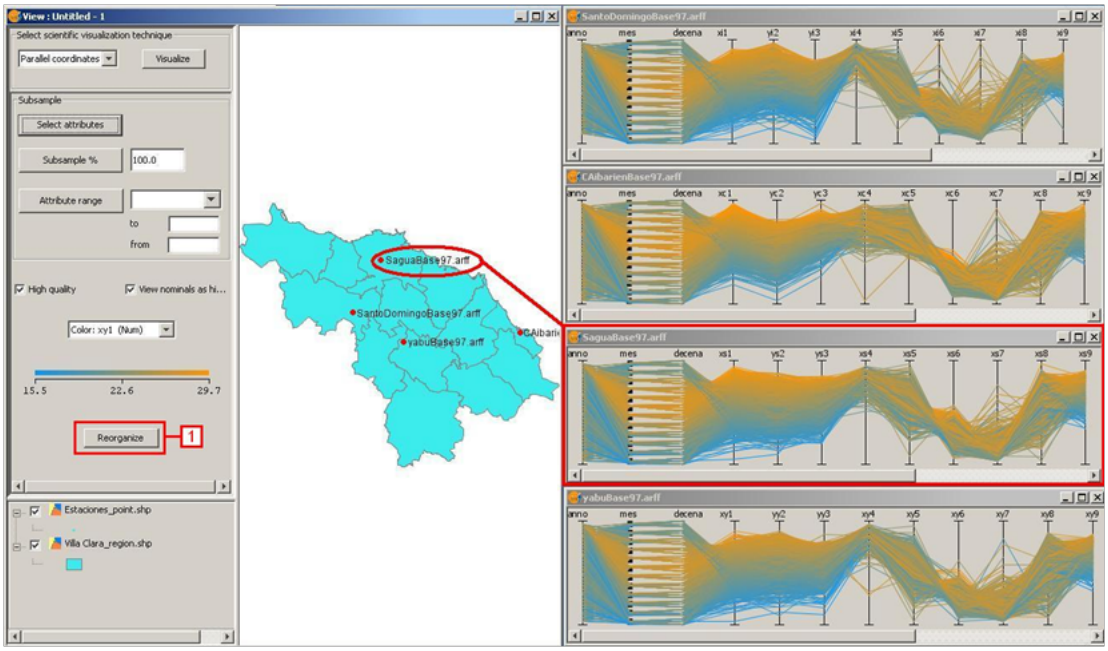

Figura A.23 Visualización coordinada utilizando coordenadas paralelas.

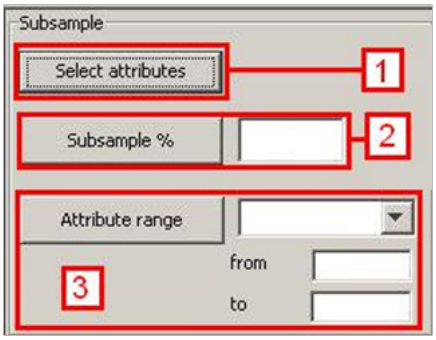

Figura A.24 Obtención de muestras del conjunto de datos.

visualizadas sobre los mapas. El componente resaltado con color amarillo muestra los mapas con los que cuenta el proyecto, además permite algunas configuraciones como son cambiar el color de fondo y la simbología de los mapas.

Coordenadas paralelas

Se mencionó anteriormente la posibilidad de que algunas técnicas fueran visualizadas sobre el mapa y otras no, teniendo en cuenta las características particulares de cada una. La técnica coordenadas paralelas, accesible desde la opción de selección de las técnicas, no es visualizada sobre el mapa. La visualización de cada conjunto de datos se realiza en paneles independientes y se agrega el nombre del archivo de datos a la localización que le corresponde en el mapa. La figura A.23 muestra un ejemplo de lo anteriormente expuesto.

Cada panel muestra la representación visual coordinada de un conjunto de datos. El nombre del archivo de datos representado se refleja en la barra de título del panel y a su vez en la localización que le corresponde en el mapa.

Un elemento común en todas las técnicas, independientemente de que la visualización sea en paneles o sobre el mapa, es la opción de obtener muestras del conjunto de datos. Obsérvese la figura A.24.

Mediante esta opción se pueden realizar las siguientes opciones:

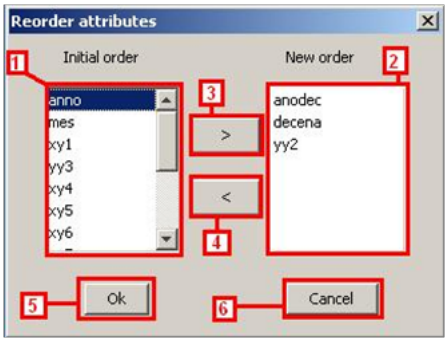

Figura A.25 Diálogo para establecer un nuevo orden de visualización de los atributos.

1. Botón *Select attributes*: Selección de los atributos de interés para la visualización.
2. Botón *Subsample %*: Obtención de una muestra para cada conjunto de datos dado el por ciento insertado en el campo de texto.
3. Botón *Attribute range*: Obtención de una muestra para cada conjunto de datos dado el rango de un atributo; el atributo se selecciona en el *combobox* y el rango de valores se inserta en los campos etiquetados *from* y *to*.

La técnica coordenadas paralelas en este tipo de visualización posee además las opciones *View nominals as histogram* y *High quality*. También permite asignar los colores a las observaciones según los valores de un atributo, esto se logra mediante un mapa de colores. En la figura A.23 con numeración uno, se muestra la opción *Reorganize*, que muestra en diálogo donde el usuario define el orden en que desea que los atributos sean visualizados (obsérvese la figura A.25).

El diálogo para establecer el orden de los atributos cuenta con una lista con el orden inicial de los atributos, identificada con el número uno; una lista con el orden final deseado, identificada con el número dos; botones para la inserción de los atributos de una lista hacia otra (tres y cuatro); botón *Ok* identificado con el número cinco para aceptar el nuevo orden establecido y el botón *Cancel* para cancelar las operaciones realizadas en esa ventana. El nuevo orden debe estar definido completamente para poder aceptar la selección, por lo tanto todos los atributos mostrados en la lista uno deben ser trasladados a la lista dos. Todas las operaciones realizadas en el panel de configuración afecta la visualización de todos los conjuntos de datos simultáneamente.

Gráfico de Andrews

La técnica gráfico de Andrews (obsérvese la figura A.26) puede ser utilizada accediendo a la opción de selección de las técnicas. Al igual que en coordenadas paralelas, las visualizaciones son realizadas en paneles independientes. Permite la obtención de muestras de los conjuntos de datos con las operaciones *Select attributes*, *Subsample %* y *Attribute range*, reorganizar los atributos que serán visualizados y asignar colores a las observaciones.

Segmentos de círculo

La técnica segmentos de círculo se encuentra disponible en la opción de selección de las técnicas de visualización. Al igual que las técnicas anteriormente expuestas permite obtener muestras de los conjuntos de datos y la reorganización de los atributos. La visualización de esta técnica se puede realizar sobre el mapa, obsérvese la figura A.27

A continuación se enumeran las opciones disponibles en el panel de configuración de la técnica:

1. *Visualization type*: Permite seleccionar cómo se realizará la visualización, es decir si es en paneles independientes o sobre el mapa. La opción seleccionada por defecto es sobre el mapa.

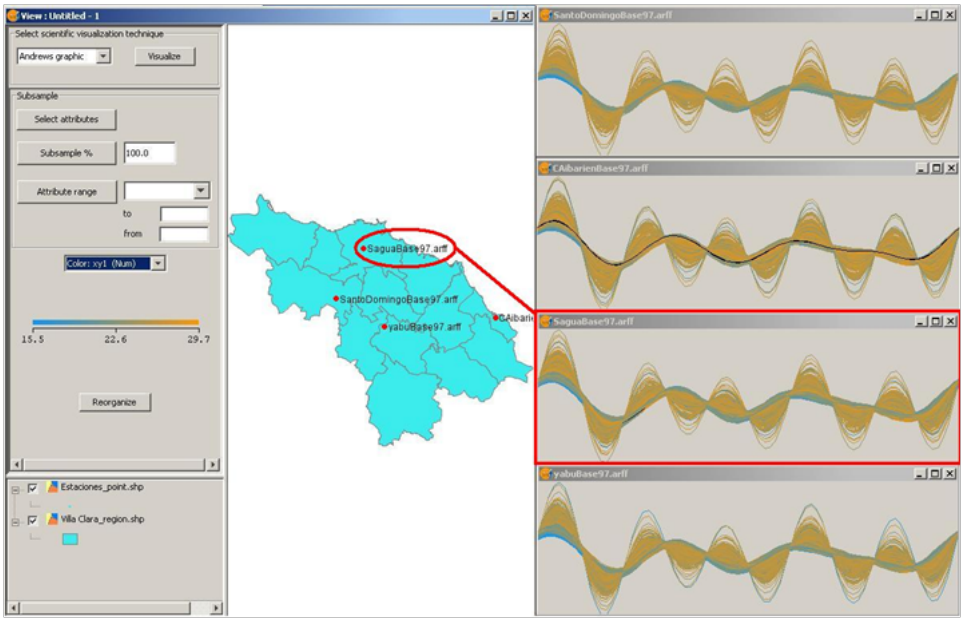

Figura A.26 Visualización coordinada utilizando gráfico de Andrews.

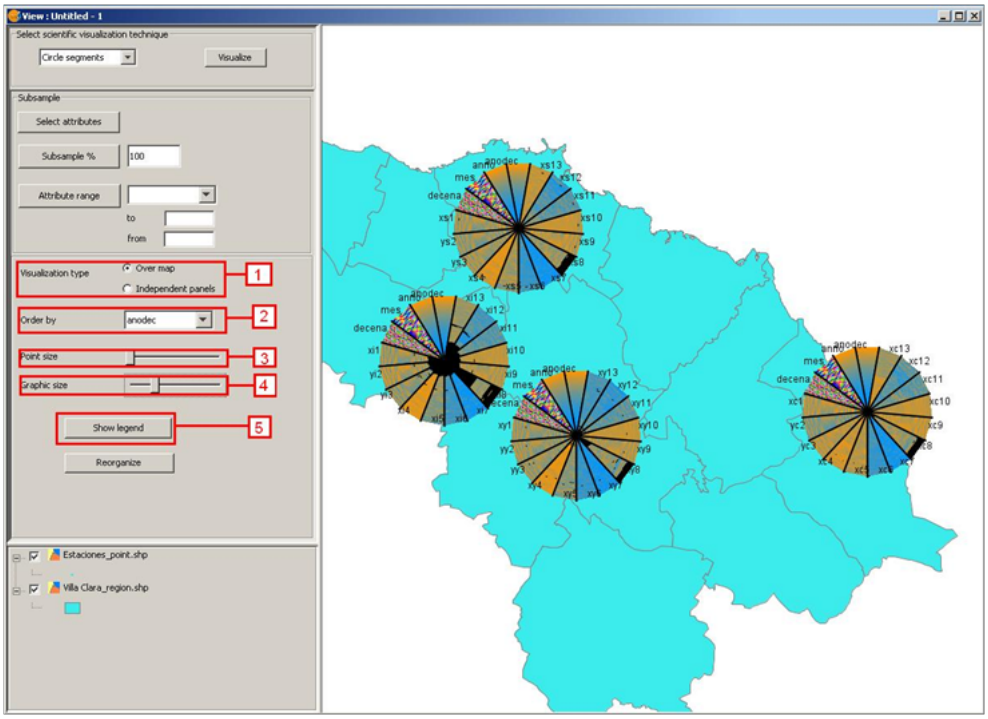

Figura A.27 Visualización coordinada utilizando la técnica segmentos de círculo.

Figura A.28 Leyenda de los valores de los atributos.

2. *Order by*: Esta opción permite que las observaciones de los conjuntos de datos sean ordenadas por los valores de un atributo que sea seleccionado.
3. *Point size*: Aumentar o disminuir el tamaño de los puntos que representan los valores de las variables para cada observación. Esta opción resulta útil cuando la cantidad de observaciones es pequeña.
4. *Graphic size*: Permite controlar el tamaño del gráfico con el objetivo de reducir el posible solapamiento de las imágenes sobre el mapa.
5. *Show legend*: Muestra un diálogo para la leyenda. La leyenda muestra la escala en colores de los valores de los atributos. ObsPermite controlar el tamaño del gráfico con el objetivo de reducir el posible solapamiento de las imágenes sobre el mapa.
6. *Show legend*: Muestra un diálogo para la leyenda. La leyenda muestra la escala en colores de los valores de los atributos. Obsérvese la figura A.28.

Patrones recursivos

Otra de las técnicas disponibles en la opción de selección de las técnicas de visualización es patrones recursivos. Esta técnica incluye las funcionalidades: visualización de muestras de observaciones de los conjuntos de datos, reorganización de atributos, aumentar y disminuir el tamaño de los gráficos generados por la técnica, ordenar las observaciones según los valores de un atributo e intercambiar entre la visualización en paneles independientes y la visualización sobre el mapa. Además permite mostrar la leyenda de los valores de las variables con el mapa de colores. Las descripciones de estas funcionalidades son similares a las explicadas en las técnicas anteriores. La utilización de esta técnica se muestra en la figura A.29 que será utilizada para exponer las opciones de interacción.

La opción identificada con el número uno permite personalizar la forma de posicionar los píxeles en la región particular de cada atributo, las posibles opciones son *Line by line* o *Back and fort*. Además el usuario puede insertar nuevos patrones mediante la opción dos de la figura. Al pulsar el botón *New pattern* se muestra el diálogo de la figura A.12 para la edición de los niveles de recursividad que conforman el nuevo patrón.

Técnicas basadas en iconos

Las técnicas basadas en iconos también están disponibles en la opción de selección para la visualización coordinada. Estas técnicas muestran un icono a la vez, según la observación que esté siendo visualizada. Si se ordenan las observaciones por una variable que denote tiempo o un orden cronológico determinado en las mediciones, el usuario puede realizar un análisis temporal del comportamiento de las variables. Los iconos implementados para este propósito son: icono en forma de estrella, icono en forma de barras y *shape coding*. La figura A.30 muestra la visualización utilizando el icono en forma de estrella, que será utilizado para explicar las opciones específicas que brinda la técnica. Un ejemplo de visualización empleando los iconos

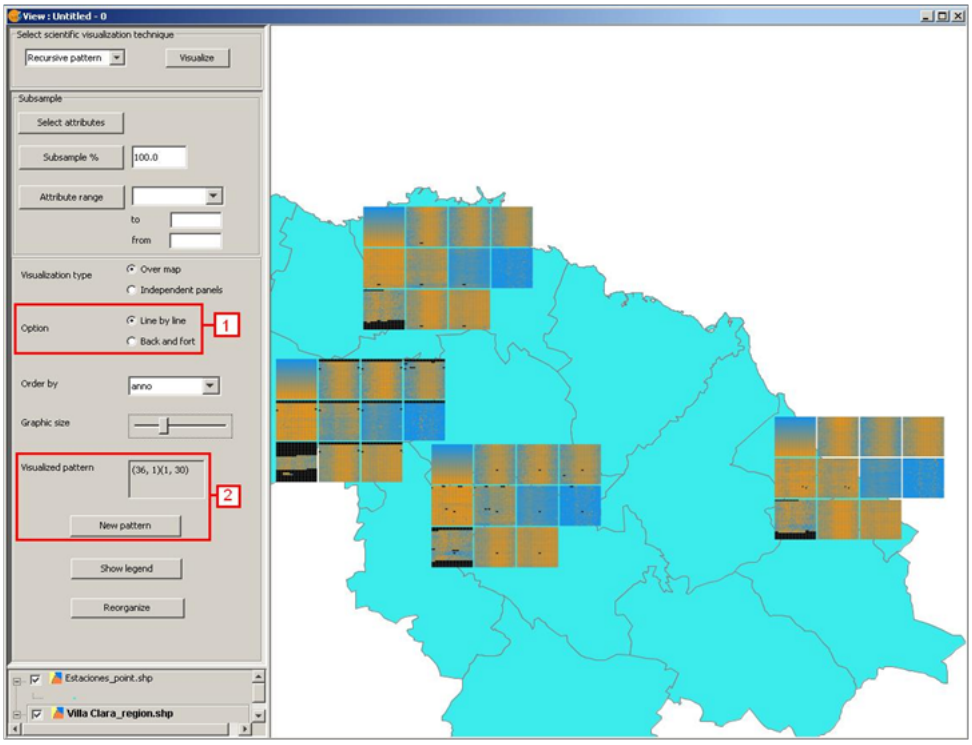

Figura A.29 Visualización coordinada utilizando la técnica patrones recursivos.

en forma de barras y *shape coding* se muestra en las figuras A.31 y A.32, respectivamente.

Estas técnicas poseen el mismo panel de configuración, por lo tanto la explicación es válida para los restantes iconos. Al igual que en las técnicas anteriores el usuario puede obtener muestras del conjunto de datos, ya sea mediante un por ciento, el rango de valores de un atributo o la selección de atributos. Puede intercambiar entre la visualización en paneles independientes y la visualización sobre el mapa, ordenar las observaciones según los valores de un atributo y reordenar los atributos.

La opción de la figura A.30 con numeración uno permite animar el proceso de visualización de los registros mediante un temporizador. Si se pulsa este botón se muestra un diálogo como el de la figura A.33 que permite iniciar, detener, acelerar y desacelerar el proceso de visualización de los registros.

La opción dos es una barra deslizante para la selección del registro que se desee visualizar, en principio realiza la misma operación que la animación pero de forma manual. La opción tres permite controlar el tamaño del icono para evitar el solapamiento de los gráficos en la imagen.

La leyenda para el caso de los iconos en forma de estrella y en forma de barras muestra la codificación en colores de los atributos, esto se utiliza para identificar cada atributo según el color que se muestra en la leyenda (obsérvese la figura A.34).

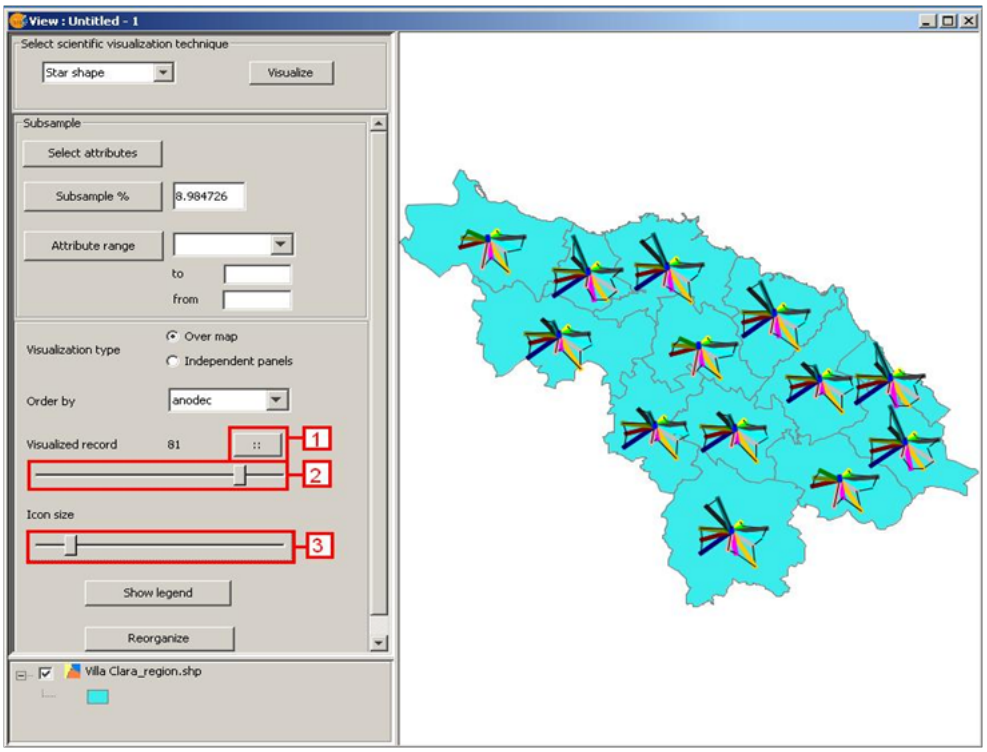

Figura A.30 Visualización coordinada mediante iconos en forma de estrella. El ejemplo muestra la visualización de los iconos localizados en los centroides de los polígonos que representan a los municipios de Villa Clara.

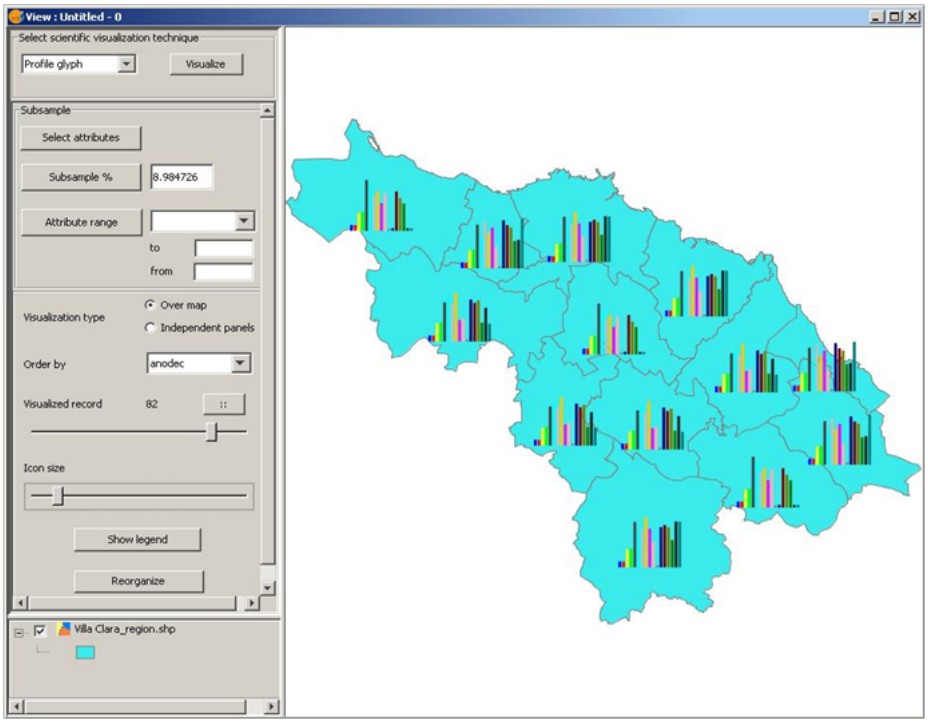

Figura A.31 Visualización coordinada utilizando iconos en forma de barras.

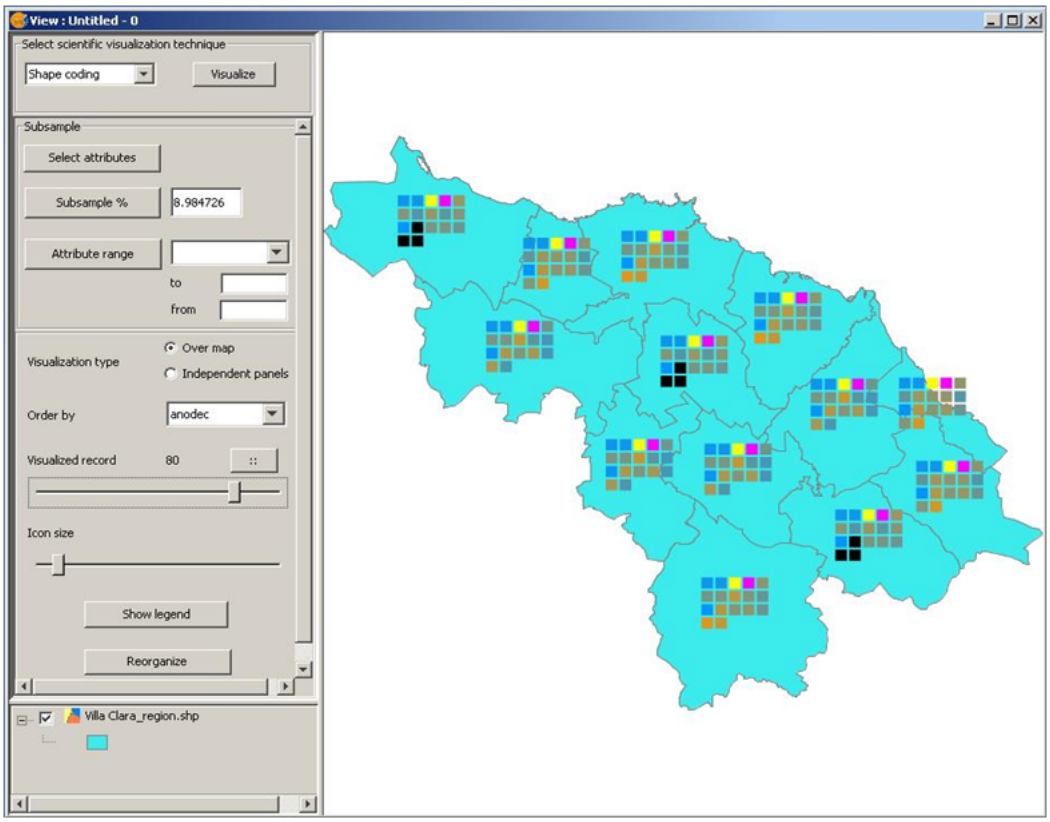

Figura A.32 Visualización coordinada utilizando *shapecoding*.

Figura A.33 Diálogo para el control de la animación.

Figura A.34 Leyenda.

Anexo 2. Ejemplos de ficheros de configuración de proyectos de visualización coordinada

El siguiente ejemplo muestra el fichero de configuración utilizado para la visualización de los datos meteorológicos de la provincia Villa Clara. En este caso se utiliza el mapa VillaClara_region.shp como mapa base y Estaciones_point.shp para las localizaciones. El campo @data muestra los archivos de datos asociados a los índices que identifican a las cuatro estaciones meteorológicas.

@map VillaClara_region.shp
@location another Estaciones_point.shp
@data
0,yabuBase97.arff
1,SaguaBase97.arff
2,CAibarienBase97.arff
3,SantoDomingoBase97.arff

En el ejemplo que se muestra a continuación las localizaciones son realizadas en el mapa base. Se utiliza el campo MUNICIPIO del mapa VillaClara_region.shp para asociar a cada municipio de la provincia un fichero de datos.

@map Villa Clara_region.shp
@locationin_map
@data
0,Corralillo.arff
1,QuemadoGuines.arff
2,Placetas.arff
3,SantoDomingo.arff
4,SagualaGrande.arff
5,Remedios.arff
6,Ranchuelo.arff
7,Manicaragua.arff
8,Encrucijada.arff
9,SantaClara.arff
10,Caibarien.arff
11,Cifuentes.arff
12,Camajuani.arff

Anexo 3. Registros informáticos

Registros informáticos en Centro Nacional de Derecho de Autor de Cuba (CENDA):

- Servicio Web para la visualización de datos multiparamétricos en GeoServer (WVS). 12/10/2010, 2194-2010
- Módulo de visualización de datos multiparamétricos para gvSIG (extSV). 12/10/2010, 2195-2010
- Aplicación Web para la visualización de datos multiparamétricos en SIG (WeVisGUI). 12/10/2010, 2196-2010.
- Extensión del módulo de datos visualización de datos espacio-temporales de gvSig (extScientificVisualization 2.0). 12/11/2012, 2499-2012
- Implementación de la operación Localización-Asignación para el módulo de análisis de redes de gvSIG (extLoc-Alloc-Network 1.0). 28/11/2013. 3268-2013
- Módulo de SEXTANTE para la manipulación de formatos de datos científicos espacio-temporales en SIGs. (extSDFSEXTANTE). 15/10/2014. 3222-10-2014
- Módulo para la visualización de datos climáticos en el Sistema de Información Geográfica gvSIG (extScientificVisualization 3.0). 15/10/2014. 3223-10-2014

Anexo 4. Encuesta principal

ENCUESTA ACERCA DE LA INTEGRACIÓN DE TÉCNICAS DE VISUALIZACIÓN CIENTÍFICA EN SISTEMAS DE INFORMACIÓN GEOGRÁFICA PARA EL ALMACENAMIENTO Y ANÁLISIS DE GRANDES VOLÚMENES DE DATOS ESPACIO-TEMPORALES
Fecha: _____ M.Sc.: ____ Dr.: ____ Esp. SIG: ____ Esp. Visualización: ____ Años de experiencia: ____

Estimado colega:

Esta encuesta pertenece al proceso de validación de la investigación: "Integración de técnicas de visualización de datos con sistemas de información geográfica para el almacenamiento y análisis de grandes conjuntos de datos espacio-temporales". Usted recibe este cuestionario porque ha sido seleccionado como Experto, por estar vinculado al uso y desarrollo de sistemas de información geográfica o al uso, diseño e implementación de herramientas de visualización. Teniendo presente su ejercicio como especialista, consideramos que su ayuda nos sería de gran utilidad. Por tal motivo, le agradecemos de antemano, que una vez que revise el material que se adjunta y que explica los propósitos de la investigación, responda la encuesta siguiente:

A partir de la necesidad de analizar las relaciones entre múltiples variables espacio-temporales georeferenciadas para solucionar un problema determinado, considere las posibilidades que brinda la integración de técnicas de visualización científica en SIG para extraer conocimiento de los datos.

1. ¿En qué grado considera Ud. que es útil visualizar múltiples variables temporales a la vez para extraer relaciones, patrones, tendencias y anomalías que están presentes en los datos?

Muy alto_____ Alto____ Neutro_____ Bajo_____ Muy Bajo____

2. ¿En qué grado considera fácil, una vez comprendida una técnica de visualización, la interpretación de los datos?

Muy alto_____ Alto____ Neutro_____ Bajo_____ Muy Bajo____

3. La manipulación de formatos de datos científicos en SIG facilita el almacenamiento y manipulación de grandes volúmenes de datos que pueden ser utilizados en visualizaciones, ¿en qué medida le resulta fácil asimilar los conocimientos mínimos para manipular formatos de datos científicos en SIG?

Muy alto_____ Alto____ Neutro_____ Bajo_____ Muy Bajo____

4. ¿En qué medida el análisis exploratorio de datos con técnicas de visualización científica facilita la comprensión inicial de la estructura y comportamiento de grandes conjuntos de datos espacio-temporales?

Muy alto_____ Alto_____ Neutro_____ Bajo_____ Muy Bajo____

5. ¿En qué medida los algoritmos para la manipulación de formatos de datos científicos en

SIG facilitan la preparación de conjuntos de datos que pueden ser analizados por herramientas de visualización?

Muy alto_____ Alto____ Neutro_____ Bajo_____ Muy Bajo____

6. ¿Qué nivel de interacción le atribuye a las técnicas de visualización del módulo de visualización científica desarrollado para gvSIG?

Muy alto_____ Alto____ Neutro_____ Bajo_____ Muy Bajo____

7. ¿En qué medida considera amigable e intuitiva la interfaz del sistema para el uso y manipulación de las técnicas de visualización?

Muy alto_____ Alto_____ Neutro_____ Bajo_____ Muy Bajo____

8. ¿Cómo valora el desempeño del sistema propuesto para el análisis de grandes volúmenes de datos? Tenga en cuenta características como eficacia, rapidez en la gestión de datos y generación de visualizaciones, consumo de recursos de cómputo, etc.

Muy alto_____ Alto____ Neutro_____ Bajo_____ Muy Bajo____

9. ¿En qué medida considera útil la integración de visualización de datos multiparamétricos en SIG para la interpretación simultánea de múltiples variables?

Muy alto_____ Alto____ Neutro_____ Bajo_____ Muy Bajo____

Anexo 5. Encuesta complementaria

ENCUESTA ACERCA DEL CONOCIMIENTO DEL TEMA DE INVESTIGACIÓN POR PARTE DE LOS ESPECIALISTAS

Estimado colega:

Le agradezco de antemano la colaboración que hace a esta investigación: "Integración de técnicas de visualización de datos con sistemas de información geográfica para el almacenamiento y análisis de grandes conjuntos de datos espacio-temporales", llenando esta encuesta.

Fecha: _____ M.Sc.: ____ Dr.: ____ Esp. SIG: ____ Esp. Visualización: ____ Años de experiencia: ____

1. El nivel de experiencia adquirido por Ud. durante su vida profesional en el trabajo con SIG es:

Alto____ Medio_____ Bajo_____

2. El nivel de experiencia adquirido por Ud. durante su vida profesional en el trabajo herramientas de visualización es:

Alto____ Medio_____ Bajo_____

3. El nivel de conocimientos que posee sobre la vinculación de SIG y Técnicas de visualización científica puede catalogarlos como:

Alto____ Medio_____ Bajo_____

4. El conocimiento actualizado que Ud. posee sobre investigaciones y/o publicaciones nacionales e internacionales sobre SIG+Visualización es:

Alto____ Medio_____ Bajo_____

5. El nivel de investigación (teórica y aplicada) que ha desarrollado en el área de SIG+Visualización puede catalogarlo de:

Alto____ Medio_____ Bajo_____

6. Con respecto a sus conocimientos generales sobre el estado de este tema de investigación, marque con una X en la raya correspondiente:

a) Los formatos de datos científicos como HDF y netCDF pueden ser útiles para manipular grandes volúmenes de datos espacio-temporales.

Totalmente de acuerdo ___ Totalmente en desacuerdo ___ Desconozco ___

b) Los SIG tradicionales incorporan los formatos de datos científicos de manera nativa.

Totalmente de acuerdo ___ Totalmente en desacuerdo ___ Desconozco ___

c) Las técnicas de visualización de datos multiparamétricos pueden ser útiles para detectar las relaciones, tendencias, anomalías y patrones en múltiples variables.

Totalmente de acuerdo ___ Totalmente en desacuerdo ___ Desconozco ___

7. El conocimiento adquirido por Ud. sobre SIG+Visualzación+Formatos de datos científicos lo ha adquirido como:

Curso postgrado_____ Maestría_____ Doctorado_____ Otro_____

Anexo 6. Publicaciones indexadas y presentaciones en eventos internacionales

Tabla A.1 Publicaciones referenciadas en importantes bases de datos.

Publicación	**Bases de datos y listados**
Vázquez-Rodríguez, Romel; Pérez-Risquet, Carlos; y Torres, Juan Carlos. 2015. Exploratory data analysis through the integration of visualization techniques in geographical information systems, *Revista Técnica de la Facultad de Ingeniería de la Universidad del Zulia*, 38(1), 73–82, ISSN:0254-0770. factor de impacto (2014): 0.047, SJR: 0.12	Science Citation Index (SCIExpanded), Compendex, Chemical Abstracts, Metal Abstracts, World Aluminium Abstracts, Mathematical Reviews, Petroleum Abstracts, Zentralblatt Für Mathematik, Current Mathematical Publications, MathSci (online database), Revencyt, Materials Information, Periódica Actualidad Iberoamericana, Journal Citation Records, Scopus
Vázquez-Rodríguez, Romel; Pérez-Risquet, Carlos; y Torres, Juan Carlos. 2013. A novel visual data mining module for the geographical information system gvsig, *Anuário do Instituto de Geociências*, 36(1), 98–111. ISSN:0101-9759. SJR: 0.25	SCOPUS (Elsevier), Web of Knowledge®, Thomson Reuters (Master Journal List), Thomson Reuters (Intelectual Property & Science), Zoological Record, GeoRef, CrossRef, EBSCO, Geoscience e-Journals
Vázquez-Rodríguez, Romel; Pérez-Risquet, Carlos; Gonzalez-Herrera, Inti.; Fajardo-Moya, Alexis. y Torres, Juan Carlos. 2010. A new visual data mining tool for gvSIG GIS. International Conference on Knowledge Discovery and Information Retrieval. ISBN: 978-989-8425-28-7	INSTICC, Scopus, DBLP

Tabla A.2 Presentaciones en eventos internacionales

Evento	**País**	**Trabajo**
Informática 2009	Cuba	Uso de sistemas de información geográfica basados en software libre para la visualización de datos meteorológicos
CompuMat 2009	Cuba	Estudio de la factibilidad de la aplicación de técnicas de visualización de datos multiparamétricos para el análisis visual de datos meteorológicos
Seminario Internacional de Doctorado en Soft Computing	Cuba	Visualización de datos meteorológicos mediante técnicas de visualización científica en SIG
Cuba-Flanders Workshop on Machine Learning and Knowledge Discovering CFWMLKD 2010	Cuba	Implementation of scientific visualization techniques for meteorological visual data in GIS
KDIR 2010	España	A new visual data mining tool for gvSIG GIS
Informática 2011	Cuba	Herramientas de visualización para el análisis visual de datos espacio-temporales en gvSIG
Geociencias 2011	Cuba	Visualización espacio-temporal mediante técnicas de visualización de datos multiparamétricos en SIG
Informática 2013	Cuba	Extensión del módulo de visualización de datos espacio-temporales de gvSIG
Workshop and Annual Meeting of the DAAD funded Network “Developing Sustainability” 2013	Indonesia	A framework for the visualization and analysis of environmental and climatic temporal sequences evenly distributed in space and time
Geociencias 2015	Cuba	Un nuevo método para el análisis visual de grandes volúmenes de datos espacio-temporales. Algoritmos para la manipulación de formatos de datos científicos en gvSIG