Implementación Eficiente de un Servicio Web de Análisis Geoestadístico Basado en Coberturas Geográficas

Fernández Herrero, C., Ríos Viqueira, J.R., Varela Pet, J., Arias Rodríguez, J.E.

Laboratorio de Sistemas, Instituto de Investigaciones Tecnológicas,
Universidad de Santiago de Compostela
Constantino Candeira S/N
tel: 981520829, fax: 981520829
chrisfh@gmail.com, joserios@usc.es, eljpet@usc.es, eljarias@usc.es

Resumen

En este artículo se presenta la implementación eficiente de las estructuras de datos, los métodos de acceso espacial y las operaciones que son el núcleo de un futuro servicio de análisis geoestadístico basado en la gestión de coberturas geográficas. El sistema combina representaciones vectoriales para coberturas de tipo discreto con representaciones raster para coberturas de tipo continuo, incluyendo métodos de acceso eficiente para cada una de ellas. Tanto las representaciones como los métodos de acceso son cuestiones de nivel de implementación alejadas de la visión del usuario, que no tiene que realizar conversiones de formato explícitas. Esto hace que, en opinión de estos autores, el sistema cumpla con el principio fundamental de independencia física de datos, bien conocido en el área de las bases de datos.

Palabras clave: Cobertura Geográfica, Análisis Geoestadístico, Métodos de acceso espacial, WCS, SIG

1 Introducción

En general, la funcionalidad proporcionada por las actuales herramientas disponibles en el área de los Sistemas de Información Geográfica (SIG) puede clasificarse en operaciones para la gestión de entidades espaciales, ciudades, ríos, carreteras, municipios, etc. y operaciones para la gestión de propiedades asociadas directamente a los puntos del espacio geográfico, generalmente resultado de medir fenómenos de naturaleza física o química. Ejemplos de estas propiedades son la temperatura, el grado de humedad, la fuerza y dirección del viento, la elevación sobre el nivel del mar, el tipo de suelo, etc. Este segundo tipo de funcionalidad se fundamenta en el concepto de *cobertura geográfica* (geographic coverage) [3] y en las operaciones definidas para su gestión [9], y es clave en muchos ámbitos de aplicación de los SIG. Un ejemplo importante de estas aplicaciones son las relacionadas con los desastres naturales como incendios, nevadas, inundaciones, terremotos, etc.

El acceso eficiente a datos geográficos ha sido también objeto de amplio estudio en SIG [1,7,8]. Dicho acceso eficiente puede estar basado en una ordenación previa de los datos mediante algún algoritmo de ordenación bidimensional o en la utilización de alguna estructura de indexación espacial. Los ordenamientos bidimensionales son en general buenas aproximaciones para el acceso a información representada en raster. Los índices espaciales, generalmente utilizados para el acceso a datos vectoriales, pueden clasificarse en dos grandes familias, según estén basados en árboles R (R-trees) o en árboles cuaternarios (quadtrees).

En la actualidad ya existen muchas herramientas que permiten la gestión de coberturas geográficas. Sin embargo, algunas de estas herramientas están especializadas únicamente en la gestión de coberturas discretas, mientras que las demás están basadas en implementaciones raster tanto de coberturas discretas como continuas, basando su funcionalidad en las operaciones propuestas por Tomlin [9]. En general, un usuario ha de ser consciente de la representación interna que tienen sus coberturas (vectorial o raster) y ha de realizar cambios en dichas representaciones para poder combinarlas en operaciones de análisis. Esto, en opinión de estos autores contradice el principio básico de independencia de datos, bien conocido en el área de las bases de datos.

Para mejorar esta situación, en este artículo se presentan estructuras de almacenamiento, métodos de acceso y algoritmos para la implementación eficiente de operaciones de análisis geoestadístico entre coberturas geográficas vectoriales y raster, que no necesitan de transformaciones de formato explícitas por parte de los usuarios. Dichos desarrollos son la base tecnológica para un servicio web de análisis geoestadístico de datos, que en opinión de estos autores, es un servicio importante de valor añadido en una Infraestructura de Datos Espaciales (IDE). En concreto, la contribución de este trabajo puede resumirse como sigue:

- Implementación de estructuras eficientes para el almacenamiento de coberturas discretas en formato vectorial y de coberturas continuas en formato *raster*.
- Implementación de métodos de acceso eficiente a las estructuras de almacenamiento anteriores.
- Desarrollo de algoritmos para la implementación eficiente de operaciones entre coberturas de cualquier tipo.
- Análisis comparativo de los métodos de acceso y algoritmos desarrollados.
- Implementación de una herramienta que permite la importación y exportación de coberturas en Geography Markup Language (GML) y la ejecución de las operaciones implementadas.

La principal ventaja de la solución propuesta respecto a las herramientas existentes en el mercado estriba en que no es necesario conocer detalles referentes a la representación interna de las coberturas para poder realizar tareas de análisis geoestadístico con ellas. Como ya se ha hecho notar, en los sistemas actuales es necesario saber si una cobertura es *raster* o *vectorial* e incluso realizar explícitamente conversiones entre estas representaciones para posteriormente poder operar. En opinión de estos autores, dichas conversiones no están relacionadas con el propósito del sistema e incrementan la complejidad en el uso del mismo.

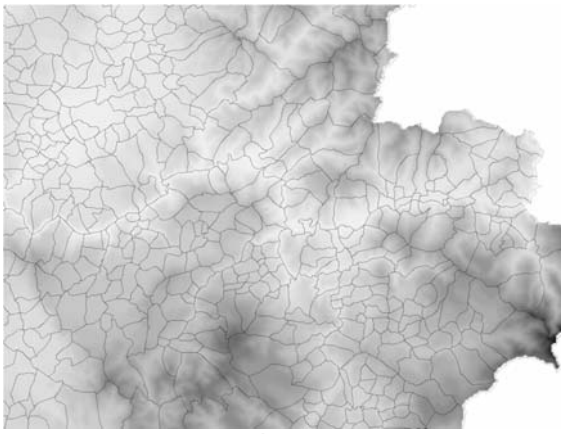
El resto de este artículo se organiza como sigue. La Sección 2 introduce de forma breve el concepto de cobertura geográfica. Detalles relacionados con la representación finita de coberturas y con su almacenamiento y acceso eficiente son el contenido de la Sección 3. La Sección 4 se ocupa de los algoritmos desarrollados y de la prueba de los mismos. Finalmente, la Sección 5 concluye el artículo y proporciona algunas líneas de posible trabajo futuro.

2 Coberturas geográficas

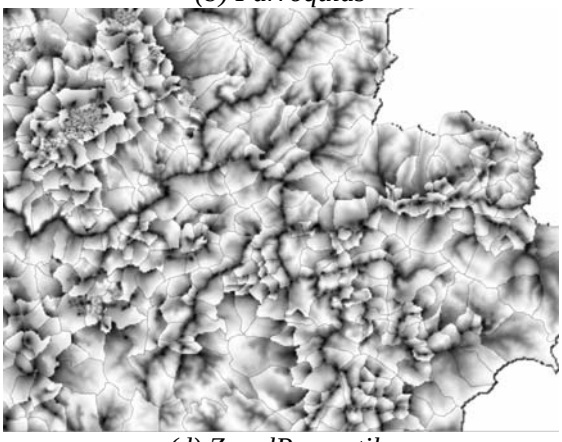
Un *Domino Geográfico* D es un subconjunto posiblemente infinito de localizaciones del espacio geográfico, es decir, de la superficie terrestre. De forma genérica, D puede tener la forma de: i) un conjunto de puntos aislados, ii) un conjunto de líneas, iii) un conjunto de superficies o iv) una combinación de puntos, líneas y superficies.

(a) Elevación

(b) Parroquias



(c) Elevación y Parroquias



(d) ZonalPercentile

Figura 1. Ilustración de coberturas y operaciones entre ellas.

Si A_i, i=1,2, ..., n denota a un conjunto de dominios alfanuméricos (enteros, reales, cadenas de caracteres, etc.) y D es un dominio geográfico, una *Cobertura Geográfica* se define como un conjunto de funciones f_i: $D \rightarrow A_i$. Cada una de las funciones f_i se llama *banda* o *atributo* de la cobertura y asocia a cada localización de dominio geográfico común el valor de una determinada propiedad del espacio.

Las coberturas geográficas pueden clasificarse con respecto a la naturaleza de su dominio geográfico en coberturas de puntos, coberturas de líneas, coberturas de superficies o coberturas híbridas. De lejos, las coberturas geográficas más utilizadas en la práctica son las de superficies. Por lo tanto, en lo que queda de artículo el término *Cobertura Geográfica* quedará restringido a este tipo de coberturas. Respecto a la naturaleza de sus atributos, una cobertura geográfica puede estar formada por atributos cuyos valores cambian en el espacio de forma continua (*coberturas continuas*) o discreta (*coberturas discretas*). La Figura 1(a) representa en tonos de gris los valores del atributo de elevación sobre el nivel de mar de una cobertura continua. Para la misma zona, el

atributo nombre de parroquia de una cobertura discreta es representado con distintos tonos de gris en la Figura 1(b). Como puede observarse en la figura, una cobertura discreta divide su dominio geográfico en *zonas* en las que los valores de sus atributos son constantes.

La funcionalidad de análisis geoestadístico incluida en la mayoría de SIG actuales se basa en las bien conocidas operaciones del álgebra de Tomlin [9], que él mismo clasifica en 4 grandes grupos: operaciones locales, operaciones zonales, operaciones focales y operaciones incrementales. Informalmente, el valor de cada localización ***p*** del resultado de una *operación local* depende del valor asociado a esa misma localización ***p*** en uno o varios atributos de entrada. En una *operación zonal*, el valor obtenido para cada localización ***p*** depende de los valores asociados a todas las localizaciones de la misma zona de ***p*** en uno o varios atributos de entrada, considerando las zonas que hayan sido definidas por otro atributo de entrada. El valor correspondiente a cada localización ***p*** en una *operación focal* depende de los valores asociados a las localizaciones del vecindario de ***p*** en uno o varios atributos de entrada, donde el vecindario de ***p*** puede definirse de distintas formas, incluyendo intervalos de distancias y ángulos. Por último, una *operación incremental* es un caso especial de operación focal en la cual el vecindario de ***p*** está restringido a las localizaciones inmediatamente contiguas a ella.

Un ejemplo de operación zonal es la operación *ZonalPercentile*. Para ilustrar esta operación consideremos de nuevo las coberturas de elevación y parroquias de las figuras 1(a) y 1(b), respectivamente, cuyas representaciones gráficas se muestran de forma solapada en la Figura 1(c). El resultado de aplicar la operación ZonalPercentile a la cobertura de elevación en cada zona definida por la cobertura de parroquias es la cobertura continua representada en la Figura 1(d). En el caso de esta operación el valor asociado a cada localización ***p*** de resultado viene dado por el cociente del número de localizaciones en la zona de ***p*** con elevación menor o igual que ***p*** entre el número total de localizaciones en la zona de ***p***, es decir, el percentil asociado a ***p*** en la distribución de elevaciones de su zona.

3 Almacenamiento y acceso eficiente

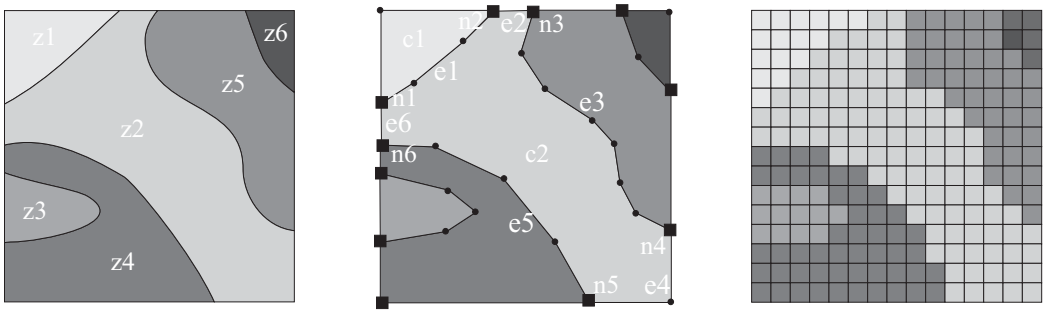
Para almacenar los datos de una cobertura geográfica de forma eficiente en una computadora se hace necesaria la utilización de alguna representación espacial finita que aproxime la naturaleza infinita de su dominio geográfico. Dicha representación finita ha de ser trasladada posteriormente a un conjunto de estructuras de datos.

3.1 Representaciones finitas

Son bien conocidas en el área de los SIG dos grandes tipos de representaciones espaciales finitas aplicadas a la representación de coberturas geográficas [6], representaciones *vectoriales topológicas* y representaciones *raster*.

Vectoriales topológicas: De forma genérica, cada una de las zonas del dominio de una cobertura se representa mediante las líneas de su borde, que a su vez son aproximadas por secuencias conectadas de segmentos, utilizando normalmente interpolación lineal en dichos segmentos. Para ilustrar esto consideremos la cobertura cuyas zonas z1, z2, ..., z6 se muestran en la Figura 2(a). Una posible representación vectorial topológica de esta cobertura es la ilustrada en la Figura 2(b). Como puede observarse en la figura, la cobertura se aproxima mediante un grafo con sus correspondientes nodos (representados mediante cuadrados en la figura), arcos (que unen pares de nodos) y caras (separadas entre sí por arcos). A esta información topológica se le une la representación geométrica que poseen tanto los nodos (puntos) como los arcos (polilíneas), que dan la forma apropiada a dicho grafo. Así

por ejemplo, la zona z2 de la cobertura se aproxima mediante la cara c2 de la representación vectorial, que a su vez está delimitada por los arcos e1, e2, ..., e6 y sus correspondientes nodos n1, n2, ..., n6. Es importante observar que la representación geométrica del arco e1 es compartida por las caras c1 y c2, y por lo tanto sólo ha de almacenarse una vez, evitando así redundancias indeseables.



(a) Cobertura (b) Vectorial (b) Raster

Figura 2. Representaciones finitas de coberturas geográficas.

Raster: En una representación raster el dominio de la cobertura se divide en superficies de forma cuadrangular y del mismo tamaño (resolución del raster) llamadas píxeles. A cada píxel se le asocia un único valor para cada atributo de la cobertura, que generaliza el conjunto de valores de todas las localizaciones contenidas en dicho píxel. Así por ejemplo, una posible representación raster de la cobertura de la Figura 2(a) es la mostrada en la Figura 2(c).

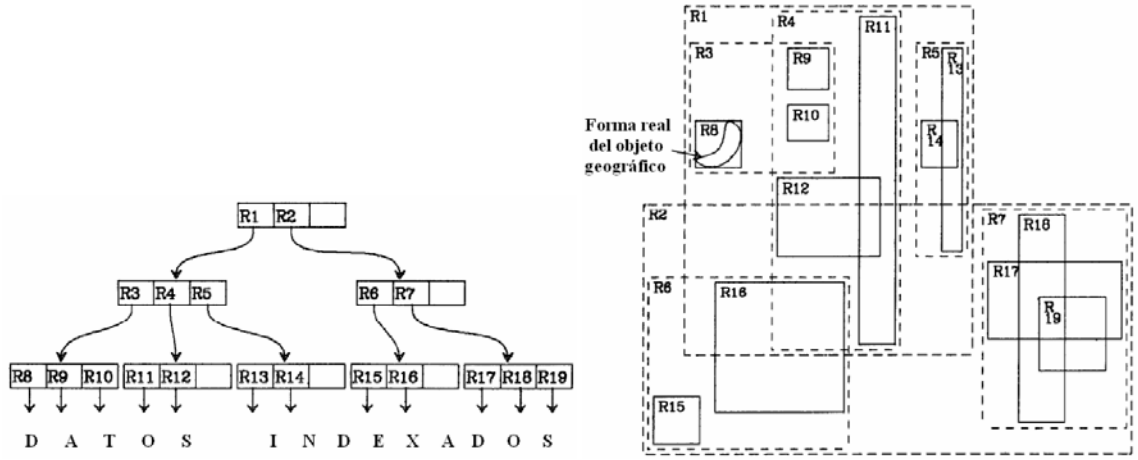
Las propiedades de las representaciones vectoriales y raster son bien conocidas en el área de los SIG [2,5,6]. Como resumen puede decirse que una representación vectorial topológica, a pesar de su complejidad lógica, permite la representación muy precisa de las zonas de coberturas geográficas discretas utilizando para ello una cantidad razonable de recursos de almacenamiento. Sin embargo, dado el gran número de zonas disjuntas distintas que por su naturaleza tienen las coberturas geográficas continuas, una representación raster es más aconsejable en este caso.

3.2 Métodos de acceso espacial

Con objeto de incrementar la eficiencia del acceso a los datos, una representación espacial finita debe ir acompañada de algún método o técnica de acceso espacial. En el contexto del presente trabajo se implementaron índices espaciales para el acceso a las coberturas vectoriales y se utilizaron curvas de relleno del espacio para la ordenación en disco de los datos de las coberturas raster.

Indexación espacial de coberturas vectoriales: Para la indexación de las coberturas vectoriales se implementaron dos tipos de índice espacial [1,6]: un árbol R (*R-Tree*) y una variante del mismo, el *STR Packed R-Tree*. Un *R-Tree* es un árbol equilibrado en altura. Cada nodo hoja contiene un conjunto de entradas, cada una de ellas formada por el rectángulo mínimo que contiene a un objeto vectorial, es decir, su MBR (*Minimum Bounding Rectangle*) y un puntero a su localización en disco. Respecto a los nodos internos, cada una de sus entradas contiene un rectángulo y un puntero a un nodo hijo, de modo que dicho rectángulo contiene a todos los rectángulos de todas las entradas de dicho nodo hijo. En la Figura 3(a) se muestra un ejemplo de un *R-Tree*, cuyos rectángulos son representados gráficamente en la Figura 3(b). Dos tipos de búsqueda son los generalmente implementados mediante *R-Tree*: búsquedas por punto (*point-query*) y búsquedas por rectángulo (*window-query*). Así, para buscar los objetos que contienen a un determinado punto P, empezando

en el nodo raíz, se busca de forma recursiva en los hijos de aquellas entradas cuyos rectángulos contienen a P. Las búsquedas por rectángulo se realizan de forma similar comprobando intersecciones entre el rectángulo R de búsqueda y los rectángulos de las entradas de los nodos del árbol. Es importante notar que, dado que los rectángulos de dos entradas hermanas pueden solaparse, la búsqueda puede continuar en paralelo por varias ramas del árbol. Este hecho reduce la eficiencia de la búsqueda, que claramente vendrá determinada en gran medida por el grado de solapamiento que existe entre los rectángulos en cada nodo. Para mejorar este comportamiento, los esfuerzos investigadores en este área han generado variantes del *R-Tree*, como son el R^+*-Tree* o el R^**-Tree*. El precio que tienen que pagar estas aproximaciones es una mayor complejidad del proceso de construcción del árbol. Una variante del *R-Tree* especialmente importante para este trabajo es el conocido como *STR Packed R-Tree*. La idea subyacente a este tipo de árbol es la ordenación previa a la inserción de todos los rectángulos de los objetos vectoriales a indexar. Una vez ordenados los objetos en base a las coordenadas de sus rectángulos mínimos, se construye el nivel de hojas del árbol de forma que el solapamiento se minimice. Los restantes niveles del árbol se construyen recursivamente a partir de las hojas. El problema de este tipo de árbol es que el conjunto completo de datos ha de ser conocido antes de la inserción, lo cual deshabilita la posibilidad de hacer inserciones y borrados posteriores a la construcción inicial. Teniendo en cuenta que una cobertura suele tratarse como un solo elemento en los niveles de aplicación, las limitaciones anteriores no resultan un problema para su indexación.

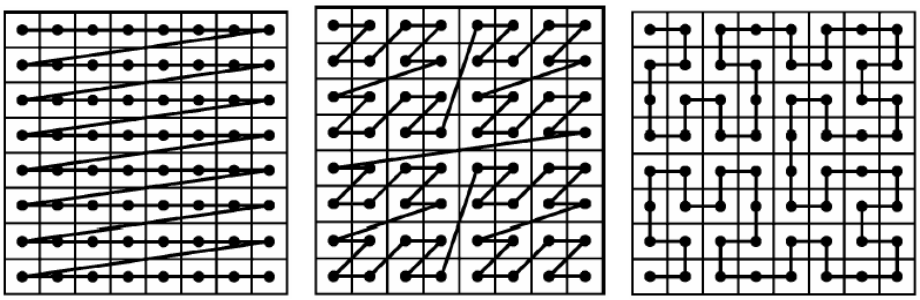


(a) R-Tree (b) Representación gráfica

Figura 3. Ejemplo de un R-Tree y la representación gráfica de sus datos.

Ordenamientos bidimensionales de datos de coberturas raster: El acceso eficiente a datos de tipo raster en el presente trabajo se basa en el ordenamiento de los valores de los píxeles en el espacio de almacenamiento. En concreto, se utilizan funciones que asignen a cada par de coordenadas (fila, columna) de cada píxel una única coordenada del espacio unidimensional de almacenamiento, de modo que los valores de los píxeles próximos entre si en 2D se almacenen también próximos. Estas funciones definen *curvas de relleno del espacio*. La Figura 4 muestra 3 ejemplos bien conocidos de curvas de relleno del espacio. Elegir el ordenamiento más apropiado para cada atributo de cada cobertura no es una tarea sencilla, ya que cada uno de ellos tiene ventajas e inconvenientes respecto a los demás. En [8] se describen algunas de las propiedades deseables para este tipo de curvas y también como cada una de ellas trata de satisfacerlas. Para realizar una búsqueda por rectángulo (*window-query*) con uno de estos ordenamientos se han de completar los siguientes pasos: 1) Las coordenadas de los píxeles contenidos dentro del rectángulo han de ser traducidas mediante el ordenamiento a coordenadas unidimensionales de almacenamiento; 2) Las coordenadas de almacenamiento se organizan en intervalos de píxeles consecutivos; 3) Los valores de los píxeles de

cada intervalo se obtienen mediante una operación de lectura. Hay que resaltar aquí que un ordenamiento que conserve la proximidad en el almacenamiento, en general incrementará la probabilidad de aparición de un menor número de intervalos de tamaño mayor, reduciendo así el número de lecturas necesarias. Además las distancias en disco entre estos intervalos serán en general menores, lo cual es una propiedad también interesante.



(a) Lineal (b) Morton (c) Peano-Hilbert

Figura 4. Ejemplos de ordenamientos bidimensionales.

3.3 Estructuras de datos

Una vez revisados los conceptos relacionados con las representaciones espaciales finitas utilizadas y con los métodos de acceso espacial implementados se describen en esta subsección de forma superficial las estructuras de datos utilizadas para almacenar tanto las coberturas vectoriales como las raster.

En general, los datos de una cobertura se reparten entre un archivo de cabecera, con formato independiente del tipo de cobertura, y varios archivos de datos, distintos para coberturas vectoriales y raster. En el archivo de cabecera se almacenan los siguientes datos.

- Un identificador del tipo de cobertura (vectorial o raster).
- El nombre de la cobertura y su descripción.
- El MBR de la cobertura, incluyendo referencia al sistema de referencia de sus coordenadas.
- Para cada uno de los atributos de la cobertura se almacenan los siguientes datos.
    - Su nombre.
    - Su tipo de dato (boolean, integer, code, measure)
    - Una tabla que asocia a cada valor de tipo cadena de caracteres el valor numérico realmente almacenado (así todos los valores almacenados son numéricos y por lo tanto de tamaño constante).
    - La referencia al espacio de códigos en caso de ser de tipo code.
    - La unidad de medida en caso de ser de tipo measure.
    - Los valores mínimo y máximo almacenados.

La descripción de los archivos de datos se proporciona a continuación tanto para coberturas vectoriales como para coberturas raster.

Almacenamiento vectorial: Además del archivo de cabecera, una cobertura vectorial necesita un archivo de polígonos, que almacena información sobre sus caras, un archivo de arcos de los polígonos y un archivo para el índice espacial. Para cada arco de la cobertura en el archivo de arcos se almacenan los siguientes datos:

- Un identificador del arco.
- El número de puntos que contiene.
- Las coordenadas de los puntos.

Para cada polígono del archivo de polígonos se almacenan los siguientes datos:

- Un identificador del polígono.
- El tamaño de la entrada del polígono en disco.
- El número de arcos del polígono.
- Para cada arco del borde exterior del polígono se almacena su identificador y su posición en el archivo de arcos.
- El número de huecos que tiene el polígono
- Para cada hueco se almacenan los siguientes datos
    - El número de arcos del hueco
    - Para cada arco del hueco se almacena su identificador y su posición en el archivo de arcos.
- Para cada atributo de la cobertura se almacena un su identificador y su valor.

Almacenamiento raster: Los datos de una cobertura raster se reparten en un archivo de domino y varios archivos de datos, uno por cada atributo. El archivo de domino almacena los siguientes datos:

- El origen de la cobertura (punto geográfico inferior izquierdo).
- Los nombres de los ejes de coordenadas del raster (generalmente X e Y).
- Las coordenadas del píxel inferior izquierdo, generalmente (0, 0).
- Las coordenadas del píxel superior derecho.
- La dimensión del raster (en este trabajo nos restringimos a rasters de dimensión 2).
- Dos vectores de desplazamiento, uno horizontal y otro vertical. Determinan la dirección de cada uno de los ejes de coordenadas y la resolución de cada píxel (Para más detalles, ver la representación de coberturas raster en GML3 [4]).

El archivo de datos de cada uno de los atributos almacena los siguientes datos:

- El tipo de dato del atributo.
- El ordenamiento bidimensional utilizado para almacenar los datos.
- El número de filas y columnas del raster. En algunos ordenamientos bidimensionales el tamaño del raster ha de ser adaptado a una potencia de 2, por esta razón estos datos no resultan redundantes con los relevantes datos del archivo de dominio.
- El conjunto de valores de los píxeles ordenados según el ordenamiento utilizado.

4 Implementación eficiente de las operaciones

En esta sección se describen de forma breve e informal los algoritmos implementados para algunas operaciones zonales y locales así como algunas pruebas de rendimiento efectuadas con los distintos métodos de indexación espacial y ordenamientos bidimensionales.

4.1 Algoritmos implementados

Como respuesta al objetivo general del proyecto, se han implementado operaciones zonales y locales entre pares de coberturas de tipo heterogéneo; es decir, una cobertura de tipo discreto representada en vectorial y una cobertura de tipo continuo representada en raster.

Operaciones zonales: Cada una de las operaciones zonales implementadas está basada de forma general en el algoritmo descrito por el siguiente pseudocódigo.

```
FOR EACH cara IN cobertura_vectorial
  mbr := obtenerMBR(cara);
  subGrid := windowQuery(mbr, cobertura_raster, atributo);
  ObtenerResultado(cara, subGrid)
END FOR
```

Como puede observarse en el algoritmo, para obtener el resultado final se procesa cada una de las caras de la cobertura vectorial. La función *obtenerMBR* obtiene el rectángulo mínimo que contiene a la cara en proceso. A continuación, dicho rectángulo se utiliza como entrada para una *window-query* sobre el raster, obteniendo como resultado un array bidimensional con los valores de los píxeles contenidos en dicho rectángulo. Finalmente, el procedimiento *ObtenerResultado* aplica el algoritmo correspondiente a la operación concreta para obtener los valores finales de la operación dentro de la cara en proceso. Estos valores pueden ser un único valor constante para toda la cara en operaciones como la media zonal (*ZonalMean*), en cuyo caso la cobertura resultante será discreta, o pueden ser valores distintos para los píxeles contenidos en la cara en operaciones como *ZonalPercentile* (ver Sección 2), en cuyo caso la cobertura resultante será continua. En cualquier caso, es evidente que el algoritmo descrito no se beneficia de la existencia de índices espaciales en la cobertura vectorial, ya que ésta se procesa completa. Por el contrario, la utilización de un ordenamiento bidimensional u otro en raster sí tiene influencia en la *window-query*. Es importante notar finalmente que en caso de aplicar la operación a sólo un pedazo de las coberturas vectorial y raster, la importancia en la eficiencia de tanto los métodos de indexación como de los ordenamientos bidimensionales sería mucho mayor.

Operaciones Locales: Cada una de las operaciones locales fue implementada mediante dos algoritmos distintos, cuyos algoritmos se describen a continuación con respectivos pseudocódigos.

Algoritmo 1

```
FOR EACH cara IN cobertura_vectorial
 mbr:=obtenerMBR(cara);
 subGrid:=windowQuery(mbr,
          cobertura_raster, atributo);
 ObtenerResultado(cara, subGrid)
END FOR
```

Algoritmo 2

```
FOR EACH pixel IN cobertura_raster
 cara:=pointQuery(pixel,
                  cobertura_vectorial);
 ObtenerResultado(cara, pixel);
END FOR
```

El pseudocódigo del algoritmo 1 es el mismo que se ha utilizado para las operaciones zonales y, como ya se ha visto, se basa en el procesado de cada una de las caras de la cobertura vectorial, por lo tanto no necesita del uso de índices espaciales. El algoritmo 2 se basa en el procesado de cada uno de los píxeles de la cobertura raster. Para cada píxel, mediante una *point-query* sobre el índice espacial de la cobertura vectorial se obtiene la cara que lo contiene. Después, utilizando la cara obtenida, se obtiene el valor resultado en el píxel en proceso. Contrariamente al algoritmo 1, el algoritmo 2 necesita de la existencia de índices espaciales para proporcionar una eficiencia razonable, siendo al mismo tiempo irrelevante el ordenamiento espacial utilizado para el raster. En cualquier caso, comparando el comportamiento de ambos algoritmos, el algoritmo 1 se comporta mejor en todos los casos.

4.2 Pruebas de rendimiento

En esta subsección se describen algunas pruebas de rendimiento realizadas a los métodos de acceso y operaciones implementadas. Para la realización de estas pruebas se consideraron dos coberturas vectoriales, cada una con 4 niveles distintos de precisión, y una cobertura raster con 10 resoluciones distintas. Más detalles de las coberturas vectoriales utilizadas se muestran en la Tabla 1.

Como puede verse en la tabla, la cobertura de municipios de Galicia MU tiene muchas menos caras

que la de parroquias PA. Sin embargo, el número de puntos en los arcos de la primera es bastante mayor en proporción. Respecto a la cobertura raster, se ha utilizado una cobertura de temperatura de Galicia a resoluciones de 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800 y 2000 metros.

Id	Caras	Arcos	Puntos	Puntos/Cara
MU0	598	1267	100793	168,55
MU1	598	1267	47829	79,98
MU2	598	1267	36397	60,86
MU3	598	1267	24736	41,36
MU4	598	1267	16421	27,46
PA0	3807	11162	63221	16,61
PA1	3807	11162	50222	13,19
PA2	3807	11162	43742	11,49
PA3	3807	11162	32416	8,51
PA4	3807	11162	25921	6,81

Tabla 1. Coberturas vectoriales (MU: Municipios de Galicia. PA: Parroquias de Galicia).

Los tiempos de construcción de los dos tipos de índice espacial utilizados se muestran en la Figura 5 en función del tamaño de nodo utilizado. Como puede verse en la figura, el *STR Packed R-Tree* se construye en menor tiempo y además su tiempo de construcción no está influenciado por el tamaño de nodo elegido. El tiempo de ordenación de la cobertura raster se muestra en la Tabla 2 en función del orden bidimensional elegido y de la resolución de la cobertura. Esta tabla muestra que existen tres ordenamientos especialmente lentos en su construcción, el *Cantor-Diagonal*, el *Peano-Hilbert* y en menor medida el *Spiral*. Esto se debe sin duda a la complejidad que tiene en estos ordenamientos el cálculo de las coordenadas unidimensionales a partir de las bidimensionales.

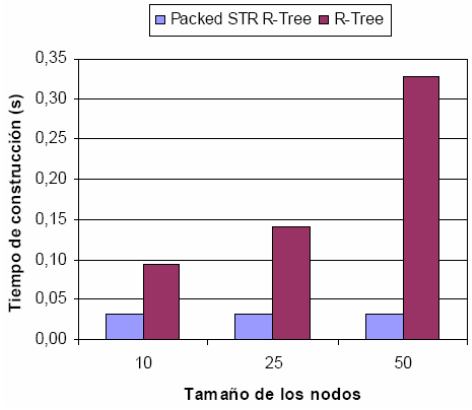
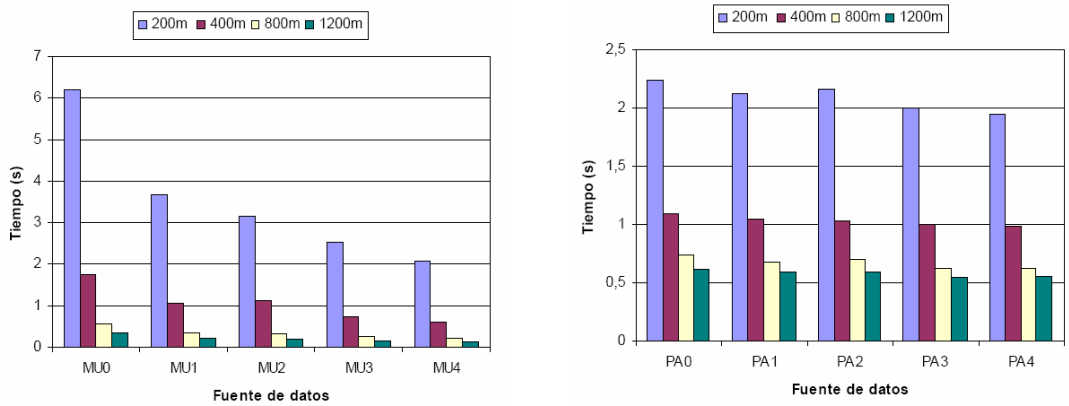


Figura 5: Tiempo de construcción de los índices espaciales

Ordenamiento	200m	400m	800m	1200m
Boustrophedonic	2,29	0,39	0,07	0,03
Cantor-Diagonal	26,98	3,53	0,50	0,07
Double-Gray	2,28	0,57	0,09	0,03
Gray	2,20	0,40	0,09	0,03
Linear	2,14	0,37	0,09	0,01
Morton	3,93	0,36	0,09	0,03
Peano-Hilbert	18,43	4,14	0,93	0,20
Spiral	6,563	0,95	0,15	0,03
U-Order	2,37	0,40	0,07	0,03

Tabla 2. Tiempo de ordenación de las coberturas raster.

En la Figura 6 se muestran los tiempos de ejecución de una operación zonal en función de la cobertura vectorial utilizada y de la resolución de la cobertura raster, siempre para un ordenamiento bidimensional fijo. En concreto, la Figura 6(a) muestra los resultados para las distintas versiones de la cobertura de municipios, mientras que la Figura 6(b) lo hace para la cobertura de parroquias. A la vista de las figuras la resolución del raster tiene una gran influencia en la respuesta del algoritmo. Respecto a la cobertura vectorial se observa que el número de puntos por cara es lo que más influye, siendo de mucha menor relevancia el número de caras de la cobertura.

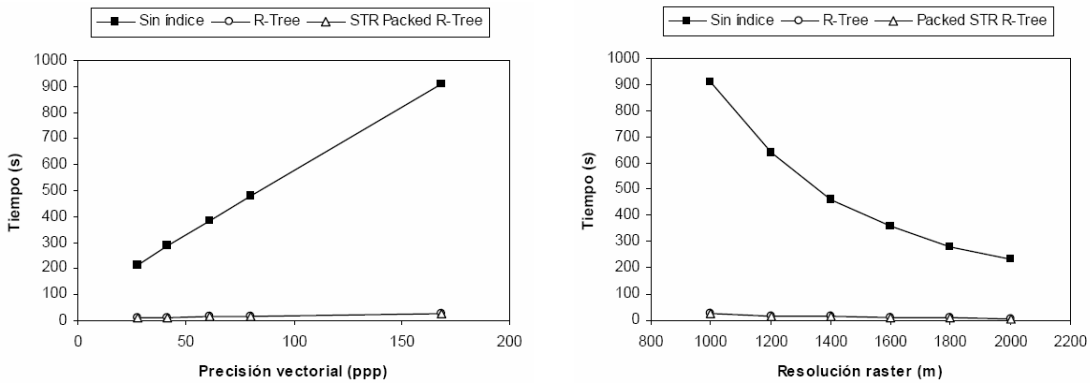


(a) Municipios (b) Parroquias

Figura 6. Tiempo de ejecución de operaciones zonales.

Respecto a la comparación de los distintos órdenes bidimensionales en las operaciones zonales se ha observado en la experiencia que los peores con diferencia son aquéllos cuyo cómputo es complejo (*Cantor-Diagonal*, *Peano-Hilbert* y *Spiral*), aunque mantengan más la localidad. Esto se debe a que las pruebas realizadas se han hecho siempre sobre coberturas completas y nunca sobre zonas pequeñas de las mismas. En pruebas adicionales realizadas sobre zonas de coberturas se ha observado que los órdenes *Boustrophedonic*, *Cantor-Diagonal*, *Linear* y *Spiral* tienen que realizar mayores desplazamientos en el disco para obtener los mismos datos. La eficiencia de los demás ordenamientos sólo se verá, por lo tanto, a la hora de procesar zonas reducidas de coberturas de gran resolución. Los resultados obtenidos para las operaciones locales implementadas con el algoritmo 1 son exactamente los mismos.

Las mejoras de eficiencia por la inclusión de índices espaciales se obtienen sólo cuando se consideran operaciones sobre partes de las coberturas vectoriales. También resultan de vital importancia para las operaciones locales implementadas con el algoritmo 2, aunque como ya se ha dicho, este algoritmo es claramente peor que el algoritmo 1. El tiempo de ejecución de este algoritmo 2 en función de la precisión de la cobertura vectorial para cada índice se muestra en la Figura 7(a). De forma similar, la Figura 7(b) muestra el tiempo de ejecución del algoritmo en función de la resolución del raster para cada índice.



(a) Precisión vectorial (b) Resolución raster

Figura 7: Tiempo de ejecución de operaciones locales (algoritmo 2)

En general la utilización de los índices espaciales se muestra como algo de gran importancia. Sin embargo, el rendimiento de los dos árboles se mueve en los mismos órdenes de magnitud.

5 Conclusiones y trabajo futuro

En este artículo se ha descrito la experiencia de implementación eficiente de algunas operaciones zonales y locales entre coberturas geográficas discretas, en formato vectorial, y continuas, en formato raster. La utilización de índices espaciales para las coberturas vectoriales y de ordenamientos bidimensionales para las coberturas raster ha demostrado ser de vital importancia para conseguir un rendimiento razonable. Las estructuras de datos, los métodos de acceso y los algoritmos implementados, junto con la experiencia de la propia implementación, son un punto de arranque interesante para la implementación de un servicio de análisis geoestadístico. Dicho servicio es un componente de interés creciente en muchos entornos de aplicación en los que los datos empiezan ya a fluir entre las distintas entidades implicadas a través de la web, es decir, futuros ámbitos de implantación de IDEs. Un ejemplo claro de estos ámbitos de aplicación es la gestión de desastres naturales, desgraciadamente foco de rallante actualidad.

En cuanto al trabajo futuro relacionado, puede mencionarse la definición de un conjunto mínimo de operaciones entre coberturas, la definición e implementación de un lenguaje de análisis geoestadístico basado en dichas operaciones, la implementación de dicha funcionalidad en un servicio web para su incorporación en una Infraestructura de Datos Espaciales (IDE), la incorporación de la coordenada temporal en el análisis y la integración con la gestión de información de entidades geográficas o features.

Referencias

[1] V. Gaede, O. Günther. *Multidimensional Access Methods*, ACM Computing Surveys, 30(2), 170-231, 1998.

[2] R. Laurini, D. Thompson, Fundamentals of Spatial Information Systems, The A.P.I.C. series Number 37, Academic Press, 1992.

[3] OGC, Topic 6: The Coverage Type and its Subtypes, version 6.0, The OpenGIS Abstract Specification, Open Geospatial Consortium, 2000.

[4] OGC, OpenGIS Geography Markup Language (GML) Encoding Specification, Version 3.0, Open Geospatial Consortium, 2002.

[5] D.J. Peuquet, “A conceptual framework and comparison of spatial data models”, D.J. Peuquet, D.F. Marble (eds.), Introductory readings in Geographic information systems, Taylor and Francis, pp. 250-285, 1990.

[6] P. Rigaux, M. Scholl, A. Voisard, Spatial Databases with application to gis. Morgan Kaufmann, 2002.

[7] H. Samet. The Design and Analysis of Spatial Data Structures, Addison-Wesley, 1990.

[8] H. Samet. Object Based and Image Based Object Representations, ACM Computing Surveys, 36(2), 159–217, 2004.

[9] C.D. Tomlin. Geographic Information Systems and Cartographic Modeling. Prentice Hall, 1990.