TEMA 3: TRANSFORMACIÓN E INTEGRACIÓN DE INFORMACIÓN GEOGRÁFICA

ÍNDICE ABREVIADO DEL TEMA

Fecha de última modificación: 02/11/20.

INTRODUCCIÓN

Las fuentes de información geográfica de las que disponemos son muy diversas:

- mapas,
- imágenes de satélite (Landstat8, por ejemplo),
- datos almacenados en dispositivos ópticos,
- datos en formatos genéricos: ODF, CSV, XML, etc.,
- obtenidos a partir de ideas propias del experto de turno.

Una vez obtenidos los datos de las fuentes, tenemos que procesarlos para obtener una ***Base de Datos con información geográfica*** almacenada. El objetivo del tema será estudiar esas operaciones intermedias de procesamiento.

En términos generales, la transformación e integración de datos consiste en

1. *Detectar y corregir errores*. Se distinguen dos situaciones:
   - si son datos con origen común, puede haber errores ocultos (pueblo con dos nombres diferentes);
   - si los datos, por el contrario, proceden de fuentes distintas, los errores suelen ser más evidentes.
2. *Resolver heterogeneidades.* Las heterogeneidades suelen dividirse en tres tipos (para datos en general):
   - técnicas (representación, formato de datos),
   - del modelo de datos (diferentes modelos de datos complican las tareas de integración de datos),
   - semánticas (datos representados con significado -semántica- diferente).
3. *Concatenación* (más específico para la información geografica). Integración de varias fuentes de datos. Hay que usar funciones para superar las diferencias entre los datos. Por ejemplo, con dos versiones de la información a tratar, hay que generar una común y expresar la incertidumbre en el resultado. Para eso es esencial el papel de los metadatos.

Nos centraremos hoy en estas fases para la información geografica.

1 DETECTAR Y CORREGIR ERRORES

Ejemplo real: si nos descargamos una capa de calles de dos fuentes distintas y las superponemos, hay variaciones, y no debería haberlas (las multilíneas señalan calles distintas). La solución no es trivial ni única. Hay que estudiar el caso.

Esas variaciones pueden ser en términos de exactitud, precisión o sesgo, o una combinación de éstas.

Algunos problemas que podemos identificar "a simple vista":

- *Exactitud posicional*. La capa coincide con la superficie real en muchas zonas, pero varía en otra. Posibles causas: por los datos categóricos (se está representando una casa, pero la capa no refleja sólo la casa, o refleja otra cosa, etc.)
- *Consistencia lógica*: coloca una casa en un sitio donde realmente no hay nada, sólo solar (por ejemplo). Es similar al anterior.
- *Completitud*: es el caso en que hay casas que existen realmente pero la capa está desfasada y nos las refleja.
- *Precisión en los atributos*. dd

En definitiva, existen cuatro principales tipos de errores:

- Errores en las coordenadas.
  - Se pueden dar *errores en el sistema de coordenadas*. Se recomienda definir un origen común (fuente) para los datos. Si no queda otra que usar varias fuentes, tratar de integrarlos al máximo y localzar errores.
  - También puede haber *diferentes unidades de medida*; en este caso, considerar la más *amplia/gruesa* (abarca más, y ahorramos espacio de almacenamiento).
  - *Datos con diferente orientación*: la solución inmediata pasa por rotarlos adecuadamente para que todos queden correctamente orientados.
  - *Confusión en las coordenadas*: (lat, lon) = (y,x), esto es un error frecuente en la info geogr que sale en los CSV. Por ejemplo, para una superficie grande, se nota mucho, o bien en sitios que no conocemos, no nos damos cuenta.
- Datos duplicados.
- Datos que faltan.
- Imprecisión o incertidumbre.

Causas de los errores:

- Si sólo se usan datos de una fuente de datos, la principal causa de los errores es que los datos sean complejos de identificar.
- Errores en la fase de transformación, al cambiar el formato, por ejemplo.

Estrategias para detectar errores:

1. Estudiar nosotros posibles inconsistencias.
2. Determinar valores imposibles.
3. Estudiar valores extremos, por ejemplo, con métodos estadísticos.
4. Comprobar inconsistencias.

Se pueden usar estrategias estadísticas. Algunas posibles:

- Muestreo.
- Análisis exploratorio de datos.
- ...

Conclusión: nos vamos a encontrar errores en los datos geográficos. Para corregirlos, usaremos diversas técnicas, más o menos profesionales (intuitivas o estadísticas, por ejemplo). Nosotros no vamos a ver ninguna técnica estadística.

2 RESOLVER HETEROGENEIDADES

Tal y como hemos comentado, en general hay tres tipos de heterogegneidades de datos de distintas fuentes:

- técnica: datos en distinto formato. No suele haber problema, teóricamente. En informacion geográfica, además tenemos que estudiar el cambio en el sistema de coordenadas.
- Del modelo de datos. Diferentes formas de almacenar y representar los datos. En el caso e info geogra, tenemos ráster y vectorial, y dentro del vectorial, hay varios. Tenemos que integrarlo todo. Algo más compleja que la primera.
- Semántica. La más compleja. Asociada al significado de los datos. Tiene q haber alguien que entida ese significado de la semántica de datos, un experto.

2.1.- Heterogeneidad técnica.

Asociada a los formatos. Nosotros elegiremos uno concreto, para raster y vectorial respectivamente, y todo lo pasaremos a esos dos formatos.

Para el caso vectorial, es más sencillo, el formato más adecuado es GeoPAckage. Es un estándar, muy compacto, de fichero único, con la codificación UTF-8, el sistema de referencia (…?).

Para el caso ráster, podemos, en principio, dejarlo en el que viene, el original.

La mayor parte de las apps GIS no soportan la conversión de datos. Usaremos herramientas de conversión intermedias que controlemos nosotros, no los que usaría la aplicación GIS, lo cual supone cierta ventaja. Conversores más frecuentes:

- GDAL/OGR,
- si es un formato más extraño o las herramientas no las entendemos bien, podemos recurrir a ciertos sitios web.

Si usamos QGIS, existe una función “on the fly”, que convierte la capa que llega al CRS común. Es muy útil, pero tiene un problema, si queremos hacer una operación de análisis sobre la capa autoconvertida, aunque sea sencilla (lo haremos en prácticas), puede que no funcione correctamente.

En definitiva, lo más recomendable es convertir las capas adecuadamente con ciertas herramientas, en la opción “Guardar como”, y ahí esojemos el CRS que queramos. Ejemplo: en QGIS.

Conclusión: si las capas tienen distinto CRS, seproyectan, con la opción “Guardar como”. Tenemos que ver la codificación.

2.1.1.- Formatos estándar para el acceso a datos.

Son aquellos que nos encontramos al acceder al información geográfica en diferentes sitios:

GeoServer, WebBrower, etc. Más frecuentes:

- OWS, el más general. Es el organismo que nos ofrece OSGeoServer (MV equivamente a nuestra imagen de prácticas).
- WCS, acceso a datos de cobertura, concretos.
- WFS: operaciones para descubrir, consultar y transofmr datos geograficos. Ofrecen operaciones de naálisis y consulta. Simple obtención.
- WMS: obtención de imagnees con las coordenadas correspondientes

2.2.- Heterogeneidad en el modelo de datos.

En el caso de bases de datos, podemos tener bases de datos con modelos de datos diferentes, y tenemos que integrarlas. En este caso tenemos raster y vectorial como principal necesidad de conversión. Si queremos hacer análisis de info geo, hay algunas que se hacen fácilmente raster-raster y vectorial-vectorial, y algunas raster-vectorial, pero hay casos en que no es fácil combinarlos en raster o vectorial. Tendremos que escoger el que más nos convenga, y no pasarlo todo a raster o vectorial por capricho. ¿Cuándo pasar a un tipo u otro? Cuando sepamos que se hará mejor en ese tipo.

Ejemplo de conversión vectorial → raster: queremos pasar multilíneas a ráster. Antes que nada, hay que definir la resolución y establecer con ello la equivalencia [longitud – celda]. Los ejemplos lo ilustran.

(imagen)

Hay que definir el criterio (a), b)), etc.

Operaciones sobre ráster:

1. Modificar la resolución: (completar)
2. Fusionar: si nos dan el raster en diversos niveles de detalle (resolución), y por páginas (cuadrados, como [MTN25, 50]), entonces tendremos que convertirlos a la misma (la más grande).
3. Recortar: como ejemplo, recortar ráster con el polígono/perímetro de nuestro municipio.
4. Clasificar en niveles: muy frecuente y útil, usando clústering para determinar polígonos y colores/degradados, etc. a cada polígono se le asocian atributos. Se usan niveles predefinidos, algunas opciones del software, etc. una aplicación directa: CORINE Land Cover.
5. Obtención de TIN (redes de triángulos). A patir del MDT, representación en forma de triángulos, que ocupa menos es más visual en determinados casos. (pg 26). idea: se define unos puntos (aleatorios, p.e.) y se definen triángulos sobre ellos con cierto criterio (hay varios, uno de ellos: triangulación de Delaunay).

Encontrar ubicaciones.- A partir de cierta información (nombre del edificio, dirección postal, localización de una ruta, etc.) necesitamos obtener puntos en el plano. Hay sitios web que hacen automáticamente esta transformación. Se trata de un problema no inmediato.

La heterogeneidad semántica se escapa de nuestros objetivos, por eso no la tratamos.

3 METADATOS

Los metadatos son datos que describen datos. A veces son minusvalorados. Recordemos que los datos en sig se dividen en geográficos, descriptivos y metadatos.

En el caso de la información geográfica, las principales necesidades que cubren los metadatos son:
- Para cada conjunto de datos:
- Además, interesa:

De esta forma, sin metadatos, se tarda mucho más en tener una visión clara del conjunto de datos (pensemos en unos datos de los que no sabemos qué se representa).

¿Qué info tiene que tener los metadatos geográficos?
- Título.
- Objetivo de su creación.
- Fuente.
- Fiabilidad. Problemas que presentan.
- ¿Cómo podemos obtener una copia?
- Estándares:
  - OGC
  - organismos federale americanos
  - ISO
- Formato de lso metadatos:
  -

Los diferentes SIG permitirán acceder de una forma u otra a los metadatos de la información geográfica de turno.

4 HERRAMIENTAS

Destacamos dos organizaciones:

- Geospatial Power Tools
  - GDAL
  - OGR
  - GDAL/OGR para R
  - PRJ.4
- GeoKettle, derivada de Kettle: permite extraer, transformar y cargar información geográfica.

5 CONCLUSIONES

1. El poder de los SIG está en combinar o integrar datos de varios tipos y analizarlos conjuntamente.
2. Usaremos datos de distintas fuentes.
3. Todas las capas deben usar el mismo Sistema de Coordenadas.
4. El uso de estándares facilita el intercambio de datos (heterogeneidad técnica).
5. Realizaremos transformaciones para resolver:
   a) Heterogeneidad del modelo de datos.
   b) Heterogeneidad semántica.

6. La calidad del resultado puede ser tan buena como la del peor conjunto usado.
7. Los metadatos favorecen el proceso de integración de datos.