Big Data & Data Science

Manejando información espacio-temporal

José Ramón Ríos Viqueira

Centro Singular de Investigación en Tecnoloxías da Información

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

citius.usc.es

Guion

- Introducción
- Datos espacio-temporales
- Captura de datos
- Análisis de datos espaciales
- Tecnologías
- Conclusiones

CiTIUS

Guion

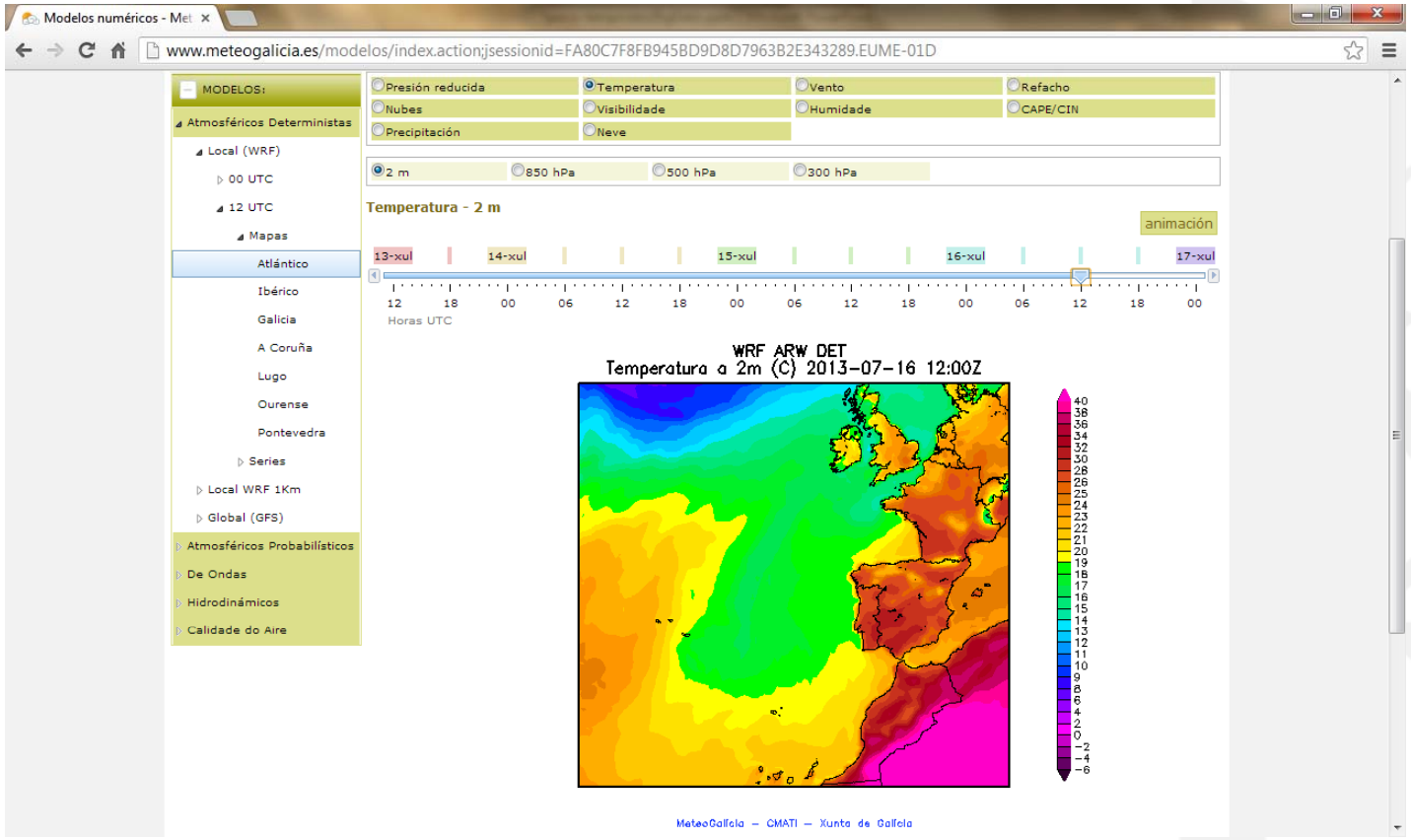
- **Introducción**
  - **Aplicaciones**
  - **Motivación**
  - **Objetivo de la charla**
- Datos espacio-temporales
- Captura de datos
- Análisis de datos espaciales
- Tecnologías
- Conclusiones

CITUS

Introducción

Aplicaciones

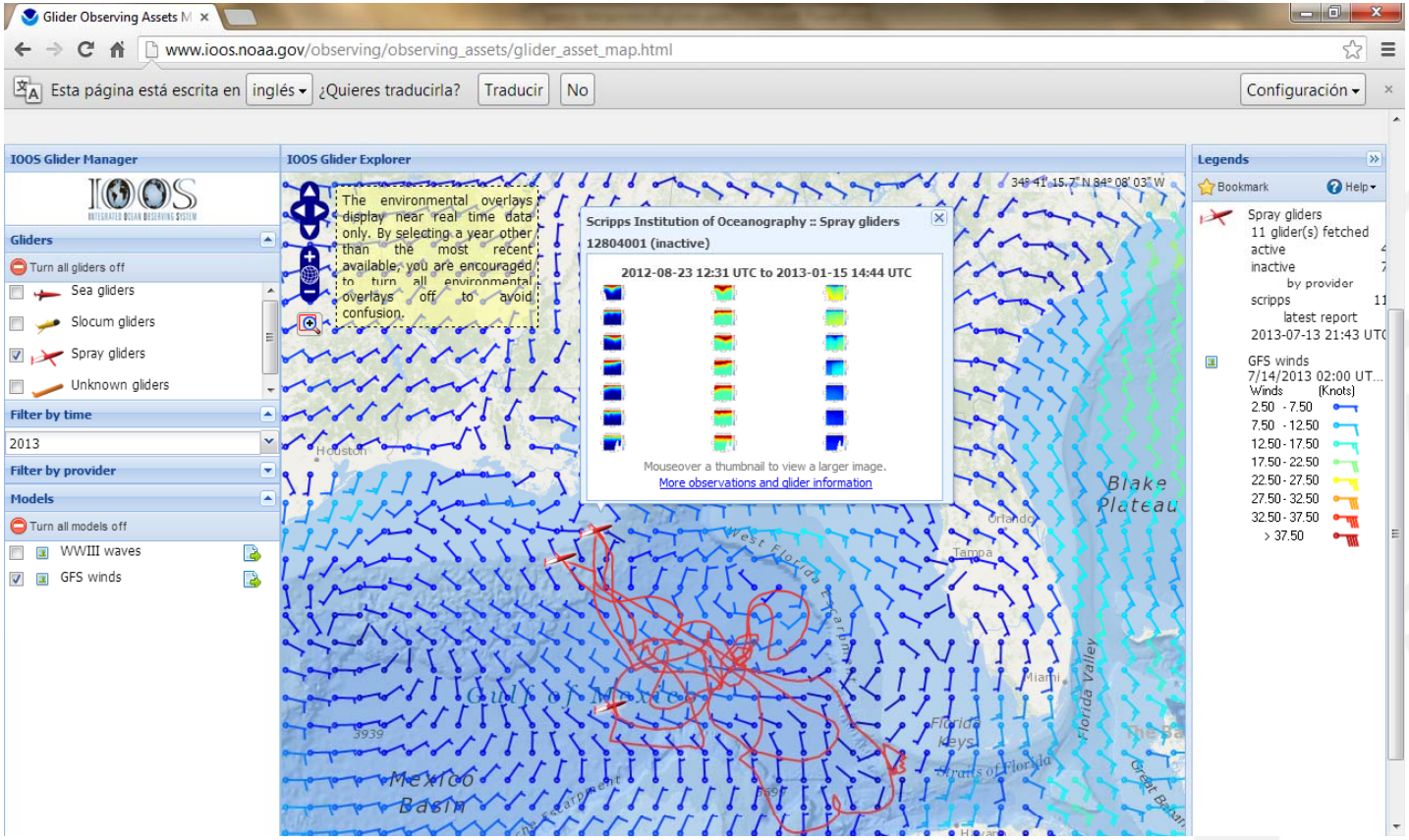
METEOROLOGÍA



Introducción

Aplicaciones

OCEANOGRAFÍA

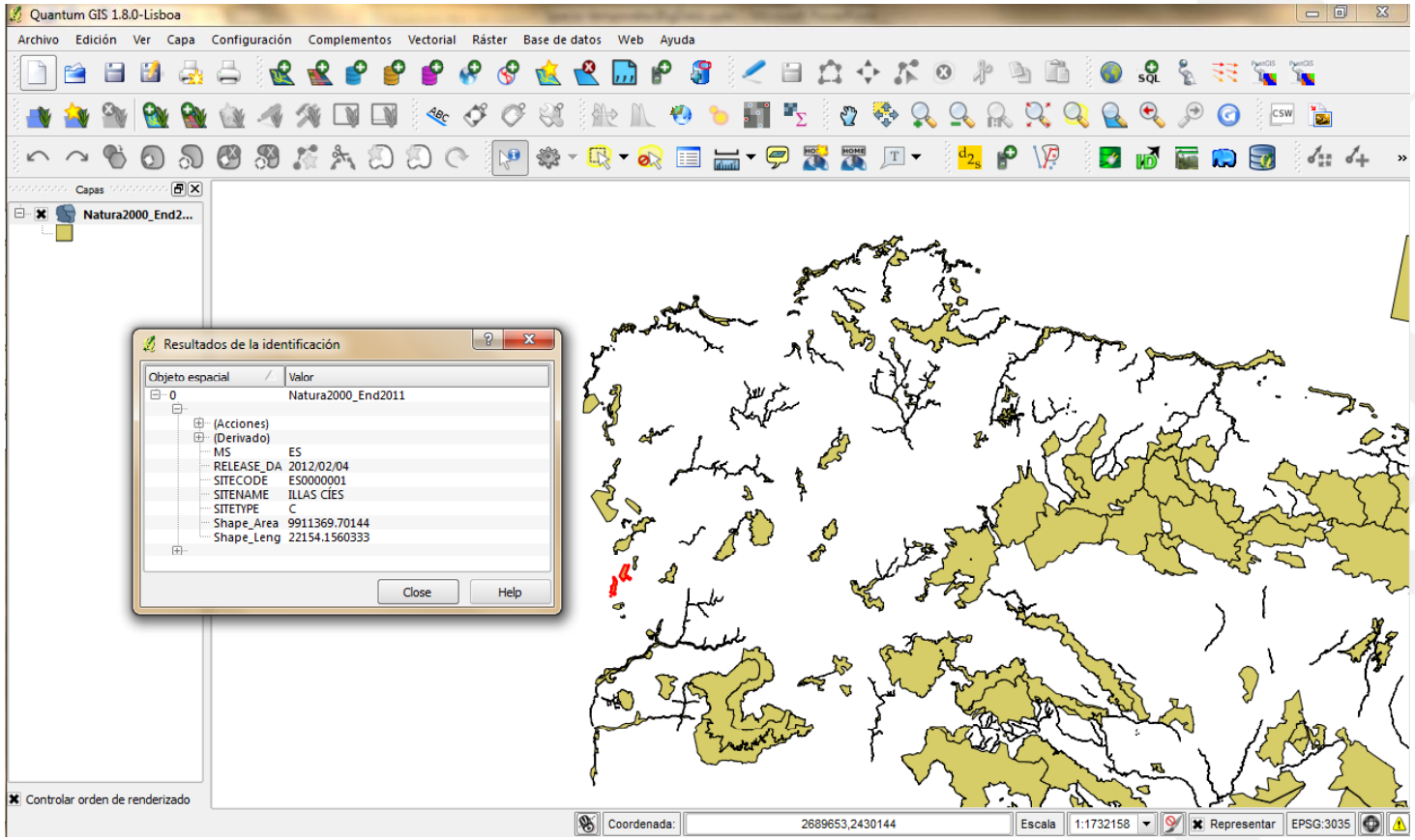


Introducción

Aplicaciones

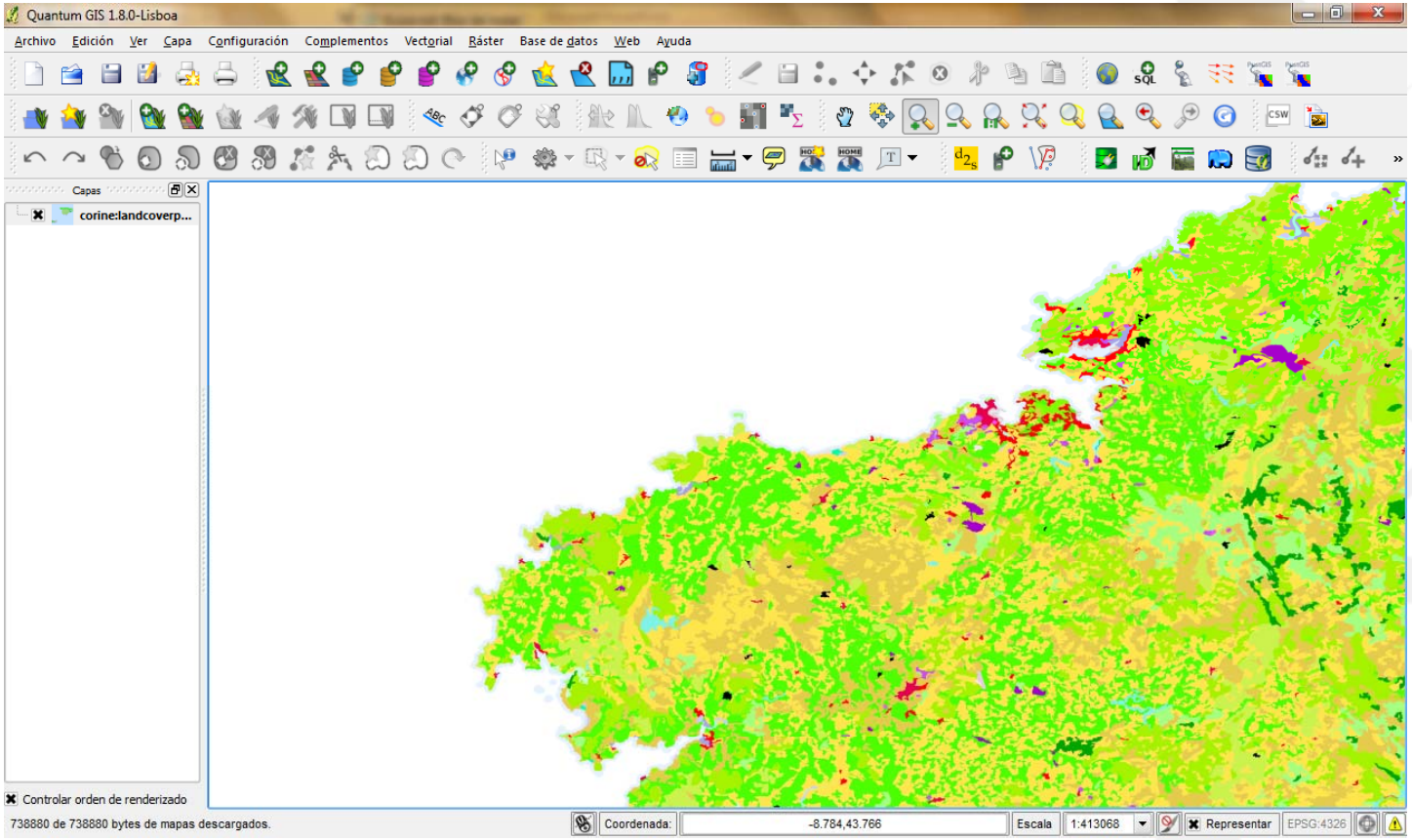
GESTIÓN MEDIOAMBIENTAL

Red Natura2000



GESTIÓN MEDIOAMBIENTAL

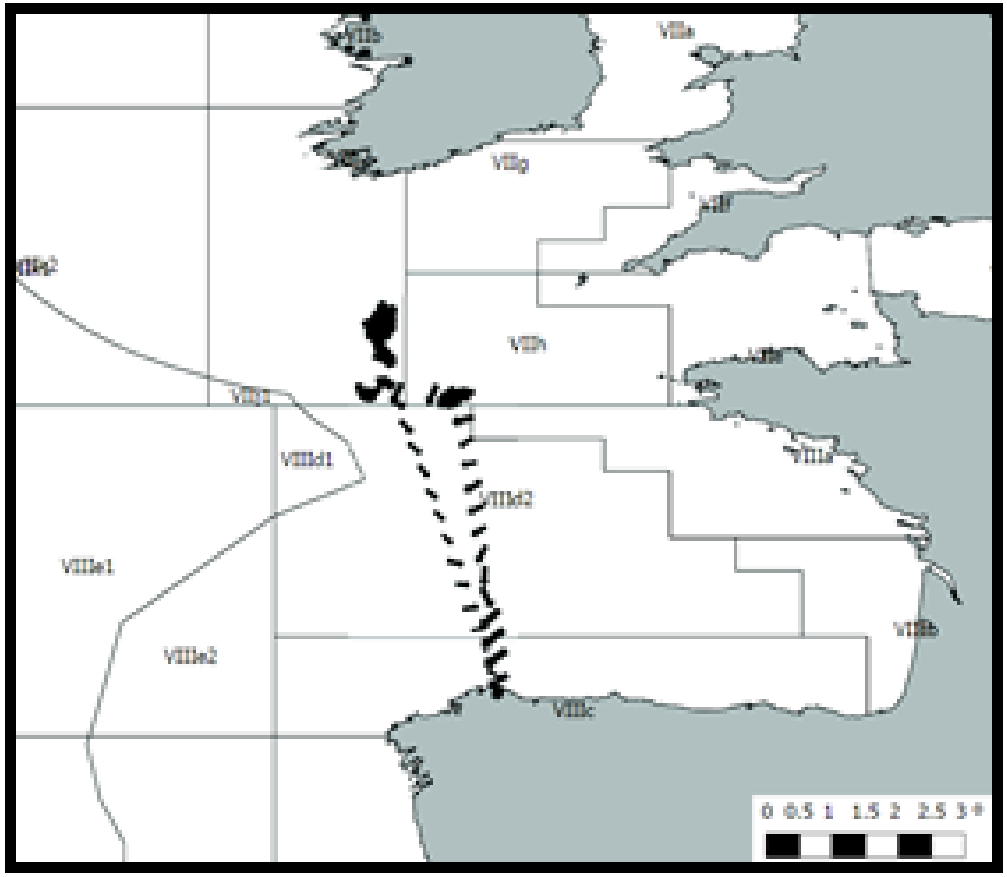
CORINE Land Cover



Introducción

Aplicaciones

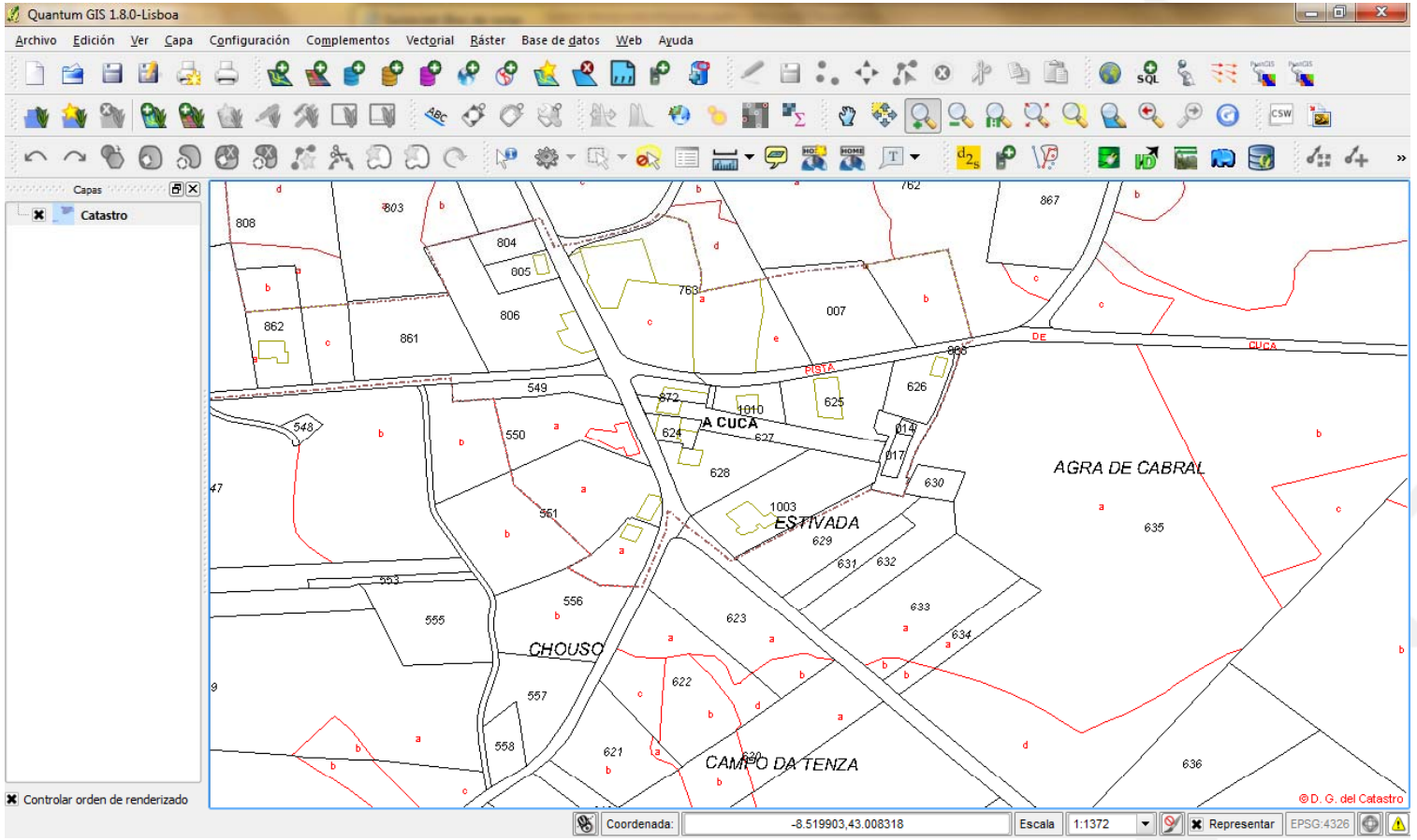
GESTIÓN DE FLOTAS



Introducción

Aplicaciones

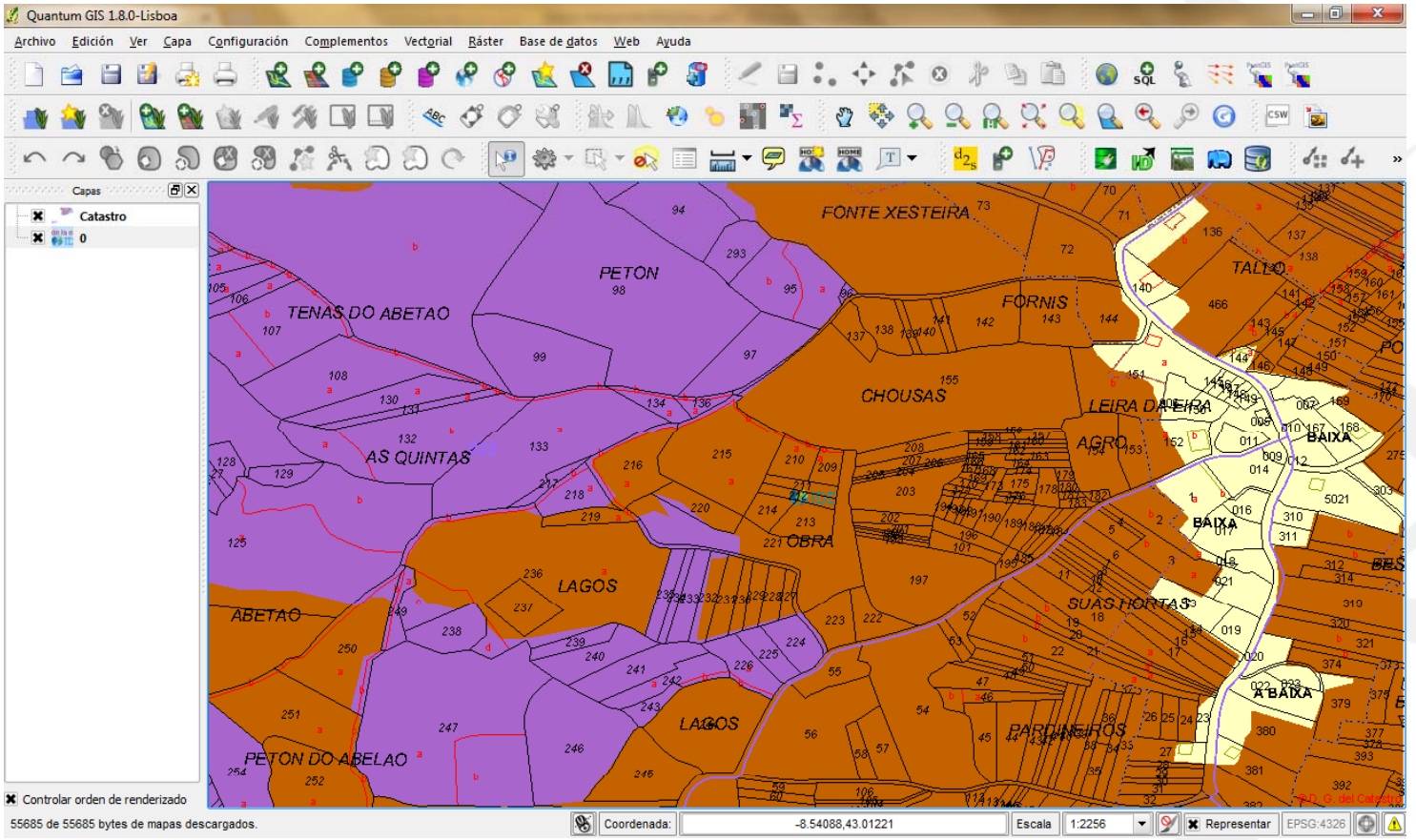
GESTIÓN CATASTRAL



Introducción

Aplicaciones

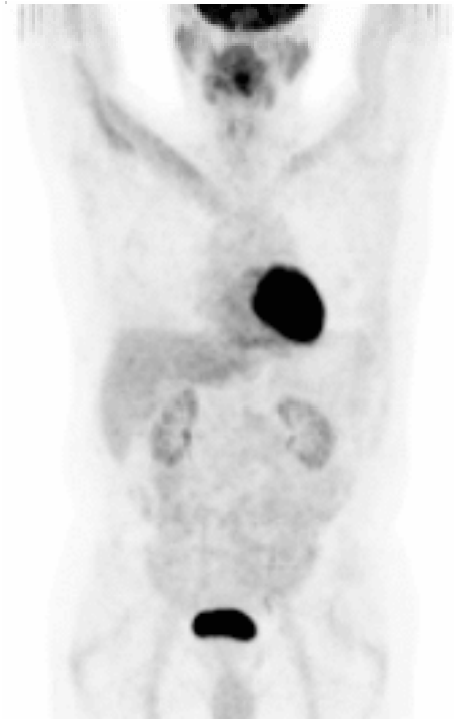
AGRICULTURA

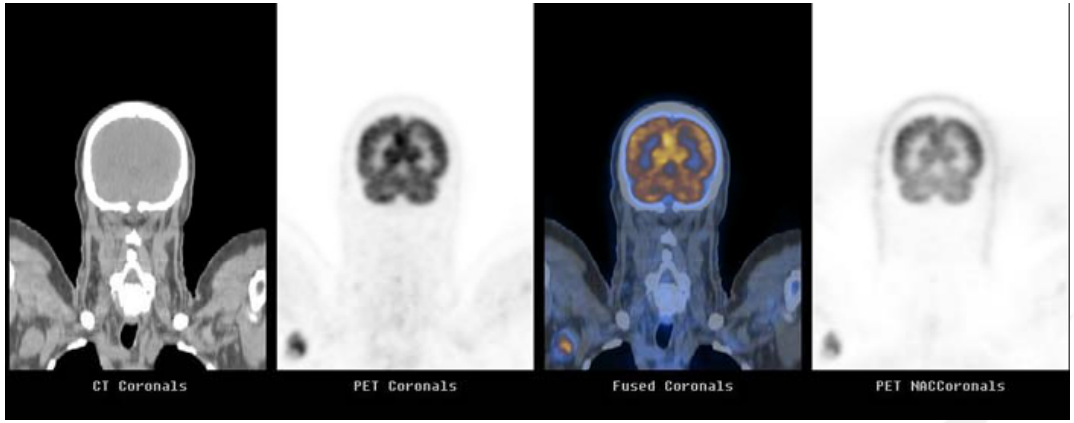


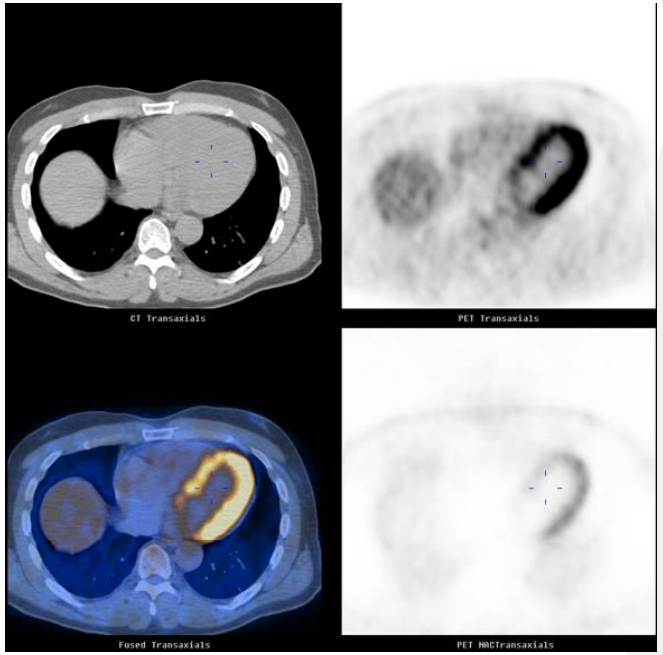
Introducción

Aplicaciones

MEDICINA





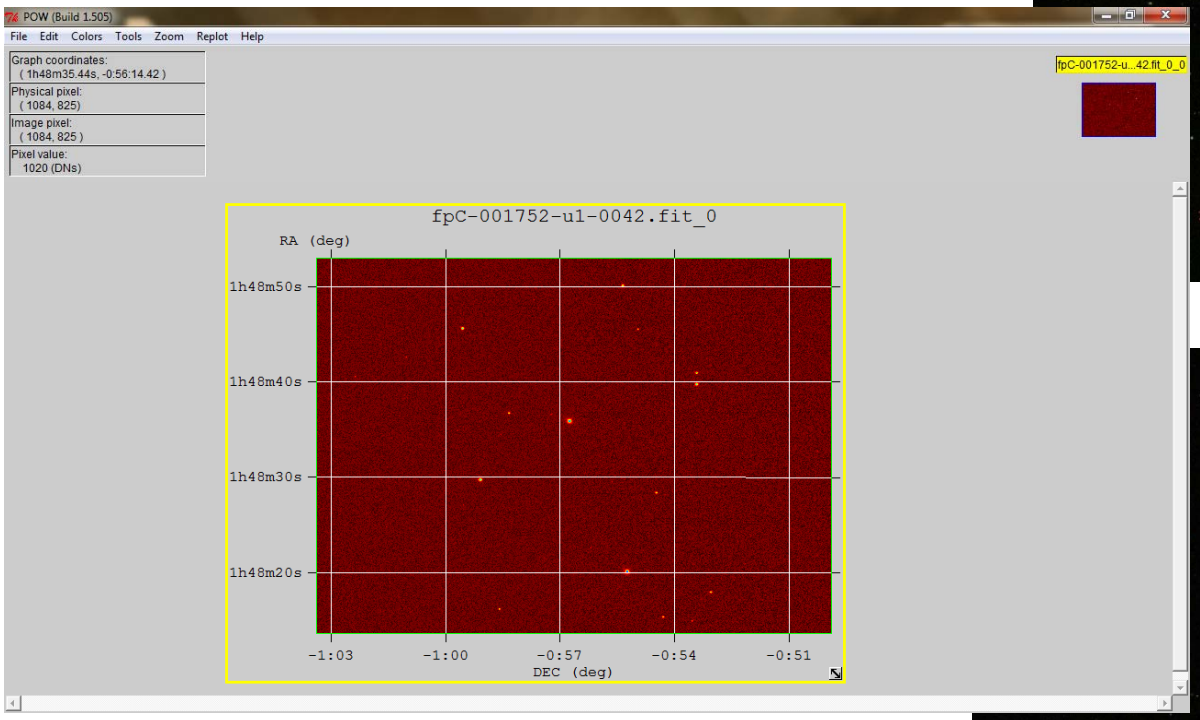


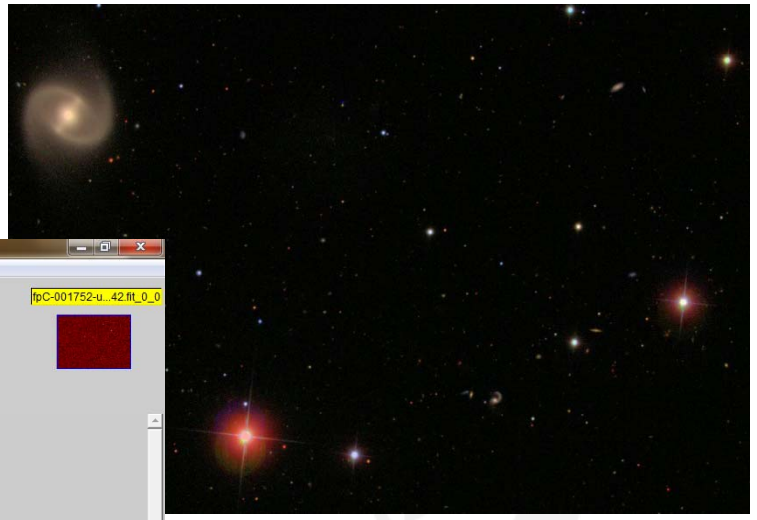
CiTIUS

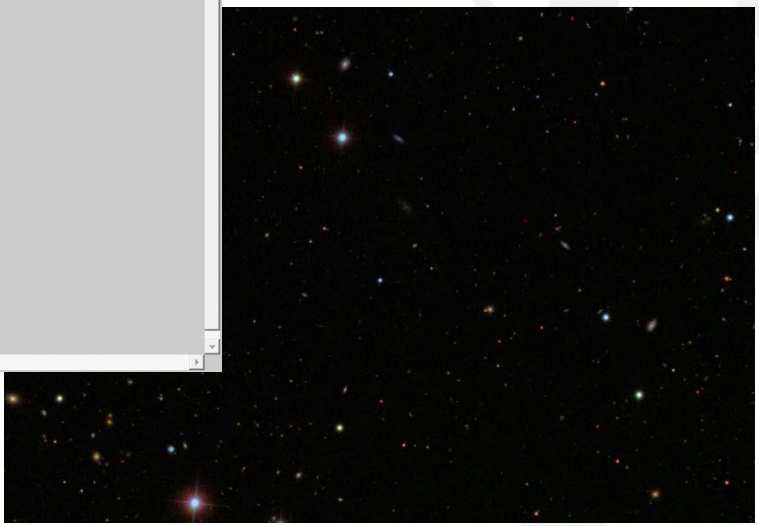
Introducción

Aplicaciones

ASTRONOMÍA







SISTEMAS DE INFORMACIÓN CONVENCIONALES

Información estructurada: Éxito rotundo del modelo relacional
Información no estructurada: Modelos de recuperación de información

Estacion

Id_est	temp	dirViento	velViento	mun
est1	15	3	46	15078
est2	22	1,23	39	15078
est4	12	0,36	25	15086

Municipio

cod_mun	nombre	pob
15078	Santiago de Compostela	96000
15086	Trazo	6000

```
SELECT m.nombre AS municipio, AVG(e.temp) AS temperatura
FROM Estacion AS e, Municipio AS m
WHERE e.mun = m.cod_mun
GROUP BY m.cod_mun, m.nombre
HAVING COUNT(e.id_est) > 1
```

DATOS TEMPORALES

Estacion

Id_est	mun
est1	15078
est2	15078
est4	15086

Medidas

Id_est	tiempo	temp	dirViento	velViento
est1	20/11/1999 15:00	15,30	1,45	22
est1	20/11/1999 15:10	15,45	1,45	22
est1	20/11/1999 15:20	16	1,56	24
...	...	...	...	....

Obtener una estimación de la temperatura a las 15:15 del 20/11/1999 en est1

```
SELECT temp
FROM Medidas m
WHERE id_est = 'est1'
      and tiempo = (SELECT MAX(tiempo)
                    FROM Medidas m1
                    WHERE id_est = 'est1'
                      and tiempo <= '20/11/1999 15:15')
```

CiTIUS

DATOS ESPACIALES

Estacion

Id_est	temp	dirViento	velViento	utmx	utmy

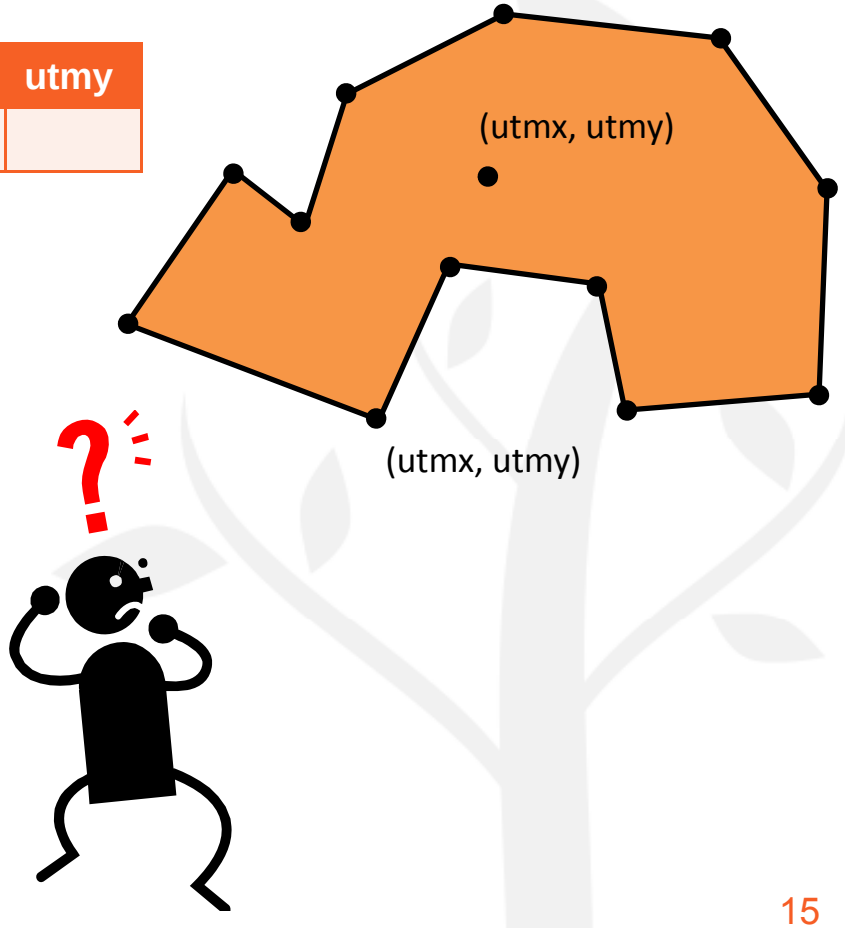
Municipio

cod_mun	nombre	pob

geo_municipio

cod_mun	num_punto	utmx	utmy

Obtener la media de temperatura en cada municipio que tenga más de dos estaciones



DATOS ESPACIO-TEMPORALES

Estacion

Id_est	utmx	utmy

Medidas

Id_est	tiempo	temp	dirViento	velViento

Municipio

cod_mun	nombre	pob

geo_municipio

cod_mun	num_punto	utmx	utmy

Utiliza el método de interpolación espacial IDW para obtener medidas de temperatura en cada punto de la provincia con una resolución espacial de 25 metros y temporal de 10 minutos

utmx	utmy	tiempo	temp

Utiliza este resultado para obtener la evolución en el tiempo de la temperatura máxima en cada municipio

CiTIUS

Introducción

Objetivo de la charla

Datos Espacio-Temporales

Breve introducción a los datos y tipos de datos espaciales y espacio-temporales

Captura y Análisis de Datos

Breve introducción a los métodos de captura de datos y a las capacidades de análisis de las herramientas y aplicaciones

Tecnologías

Breve introducción a las tecnologías de gestión de datos espacio-temporales existentes en el mercado. Ejemplos de algunas herramientas.

Guion

- Introducción
- **Datos espacio-temporales**
  - **Datos espaciales**
  - **Datos temporales**
  - **Objetos móviles**
- Captura de datos
- Análisis de datos espaciales
- Tecnologías
- Conclusiones

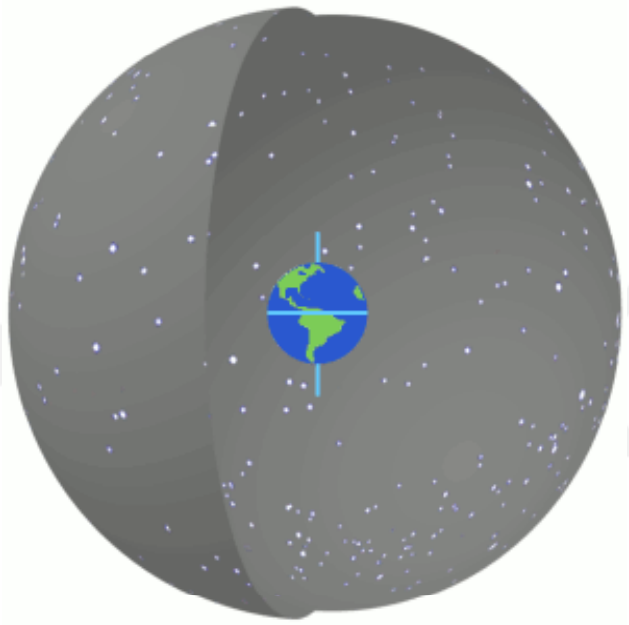
CiTIUS

Datos Espacio-temporales

Datos Espaciales

REPRESENTACIÓN DEL ESPACIO

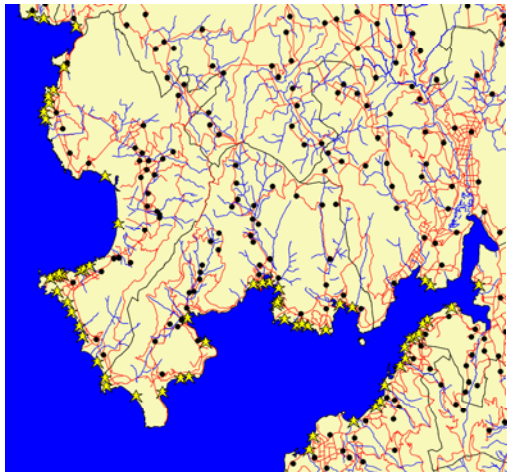
Asignar coordenadas numéricas a cada punto del espacio

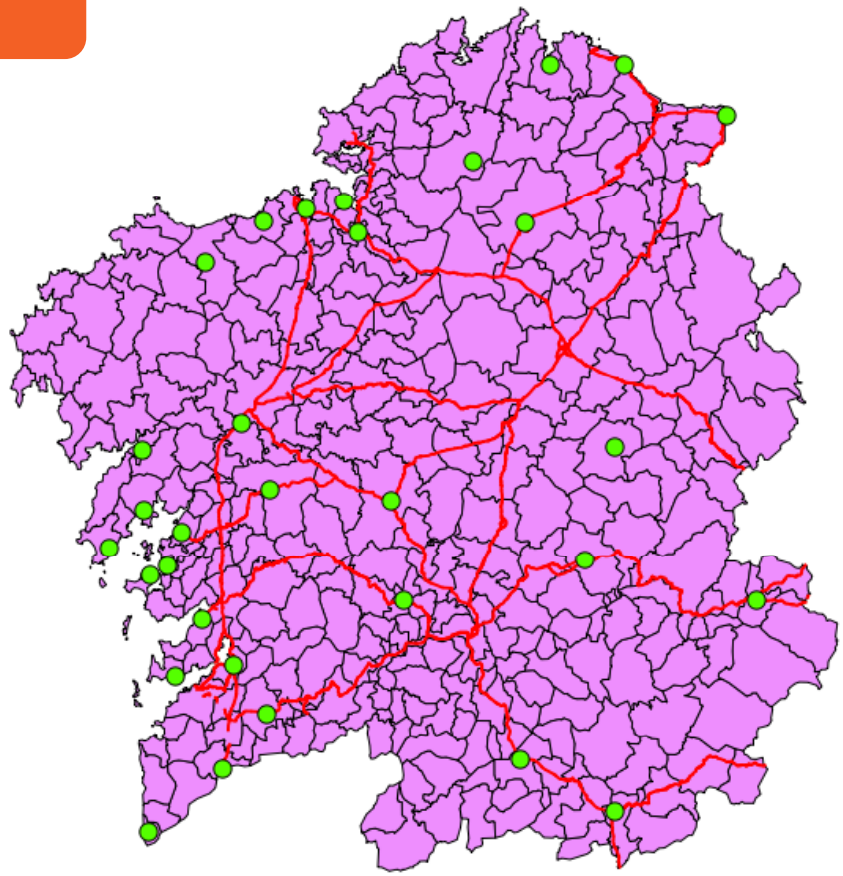


CiTIUS

TIPOS DE INFORMACIÓN ESPACIAL

- **Entidades Espaciales**
  - Propiedades convencionales
  - Propiedades geométricas
    - Punto
    - Línea
    - Superficie
    - etc.



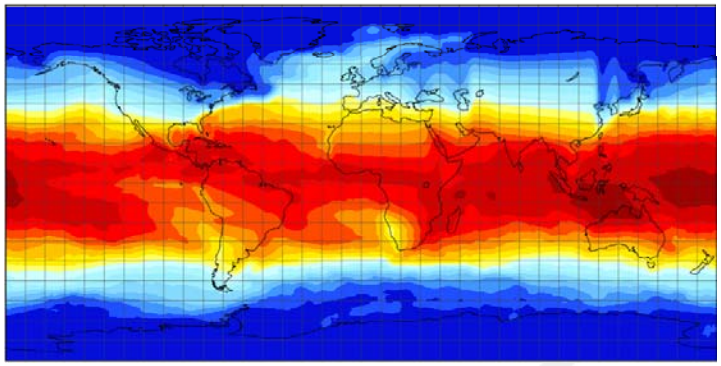


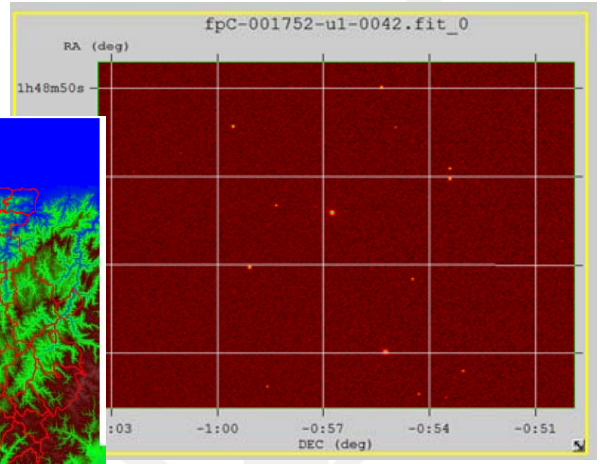
Datos Espacio-temporales

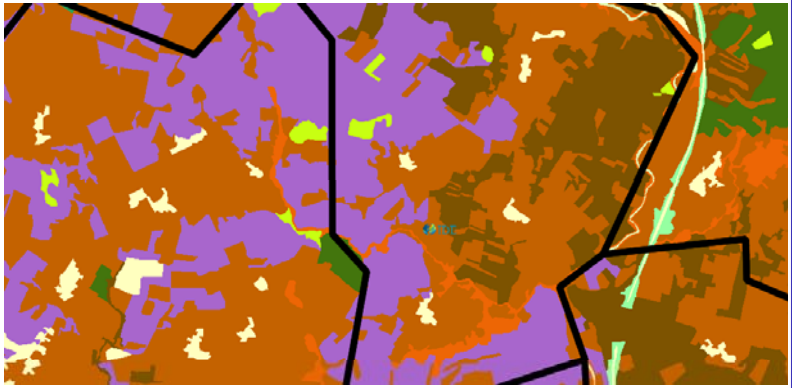
Datos Espaciales

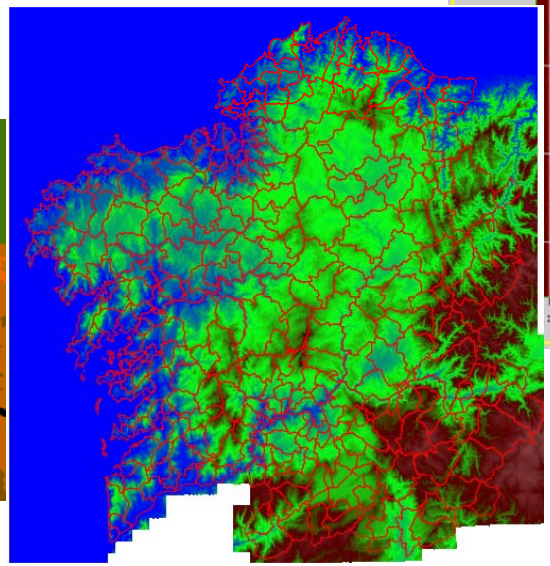
TIPOS DE INFORMACIÓN ESPACIAL

- **Coberturas Espaciales**
  - Conjuntos de funciones
  - Dominio Espacial
  - Rangos Convencionales
  - Tipos
    - Cambio continuo
    - Cambio discreto





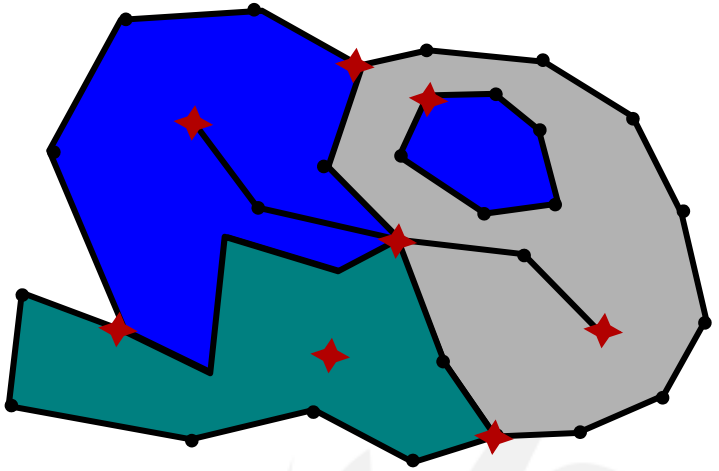


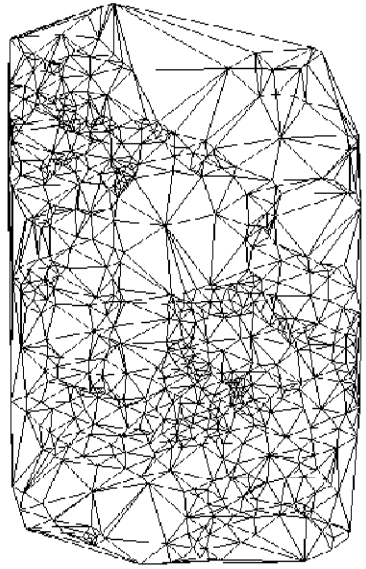


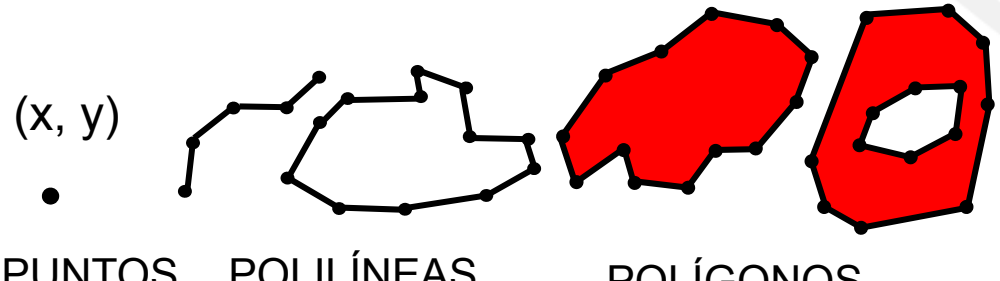
CiTIUS

Representaciones Finitas

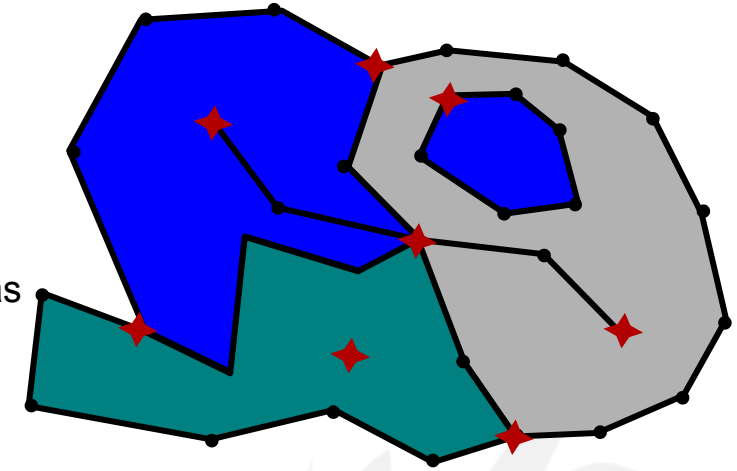
- Representaciones Vectoriales
  - **Vectorial Geométrica**
    - Puntos
    - Polilíneas
    - Polígonos
  - **Vectorial Topológica**
    - Nodos
    - Arcos (geometría polilínea)
    - Caras
    - Caso especial
      - Red Irregular de Triángulos (TIN)



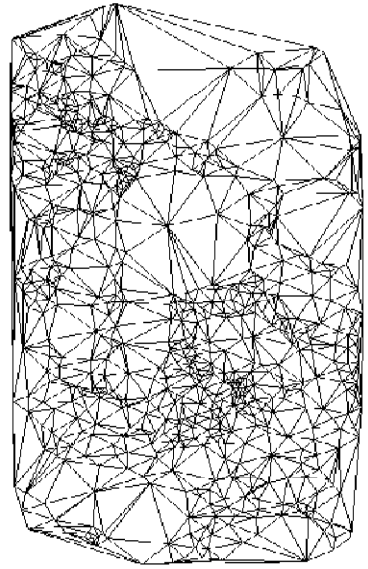


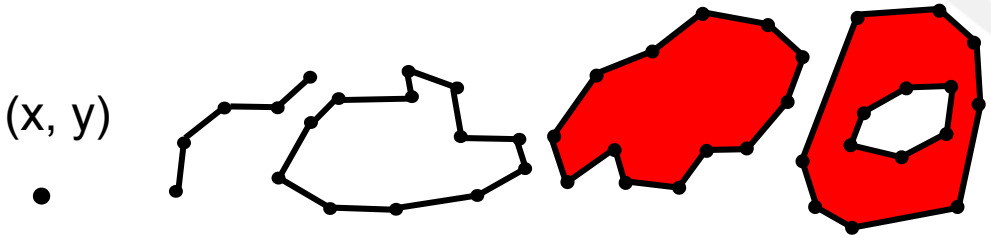


REPRESENTACIONES FINITAS



- Representaciones Vectoriales
  - **Vectorial Geométrica**
    - Ventaja: Representación compacta de geometrías
    - Inconvenientes
      - Redundancia
      - Análisis topológico poco eficiente
  - **Vectorial Topológica**
    - Ventaja: Sin redundancia. Análisis topológico eficiente
    - Inconvenientes
      - Reconstrucción topológica en inserción y borrado
      - Reconstrucción de objetos para visualización





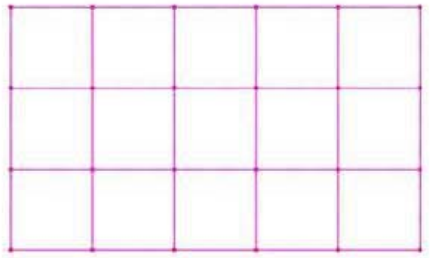
PUNTOS POLILÍNEAS POLÍGONOS

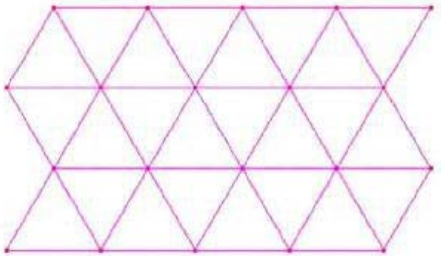
Datos Espacio-temporales

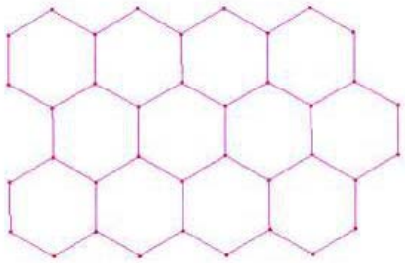
Datos Espaciales

REPRESENTACIONES FINITAS

- Mosaicos regulares
  - Triángulos, Hexágonos o Cuadrados
  - **Raster**
    - Celdas de forma cuadrada (píxeles)
- Usos comunes
  - Geometrías de entidades espaciales
    - Vectorial Geométrica
  - Coberturas discretas
    - Vectorial topológica
  - Coberturas continuas y discretas
    - Raster
  - Renderización 3D de coberturas continuas
    - TIN







CiTIUS

Datos Espacio-temporales

Datos Temporales

Tiempo de Validez

- Tiempo durante el cual un hecho es cierto en la realidad que se está modelando
- Instante, período o conjunto de períodos
- Proporcionado por el usuario

Tiempo de Transacción

- Tiempo durante el cual un hecho está almacenado en el sistema
- Período
- Actualizado por el sistema en inserciones, modificaciones y borrados

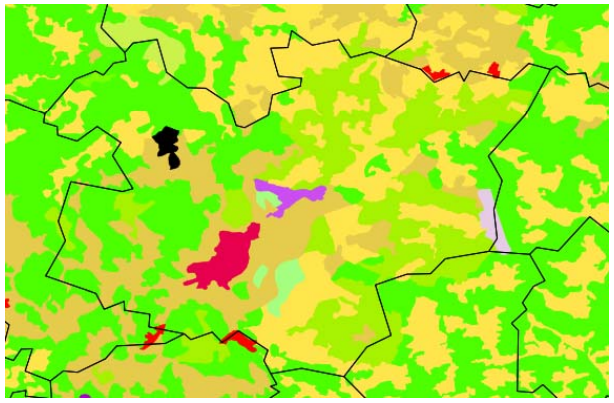
CiTIUS

Datos Espacio-temporales

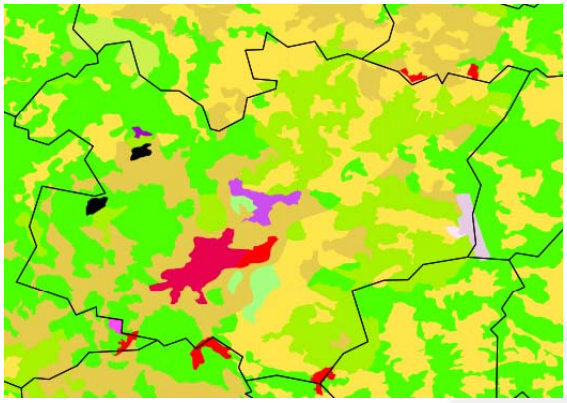
Datos Temporales

Cambio Discreto

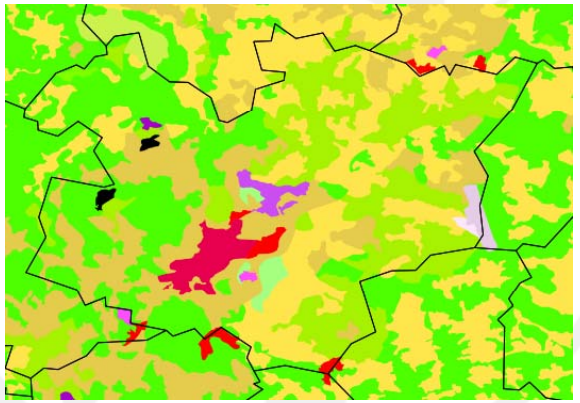
[1990, 2000)



[2000, 2006)



[2006, -)

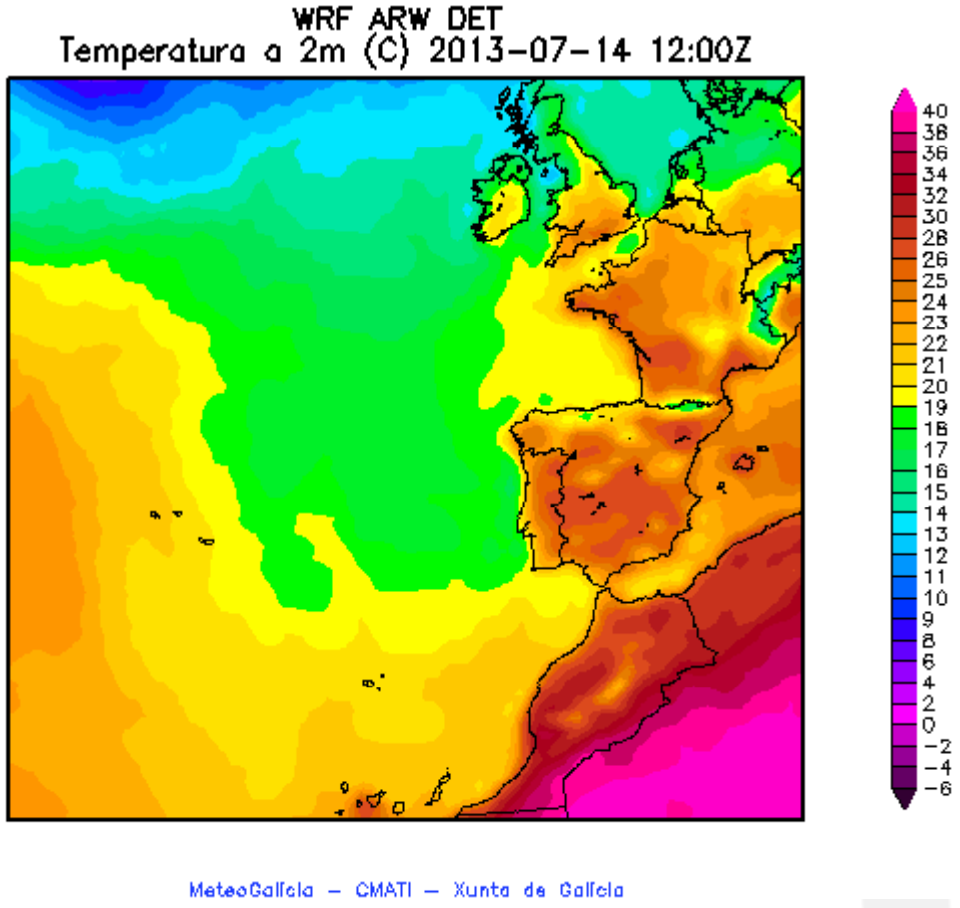


CiTIUS

Datos Espacio-temporales

Datos Temporales

CAMBIO CONTINUO

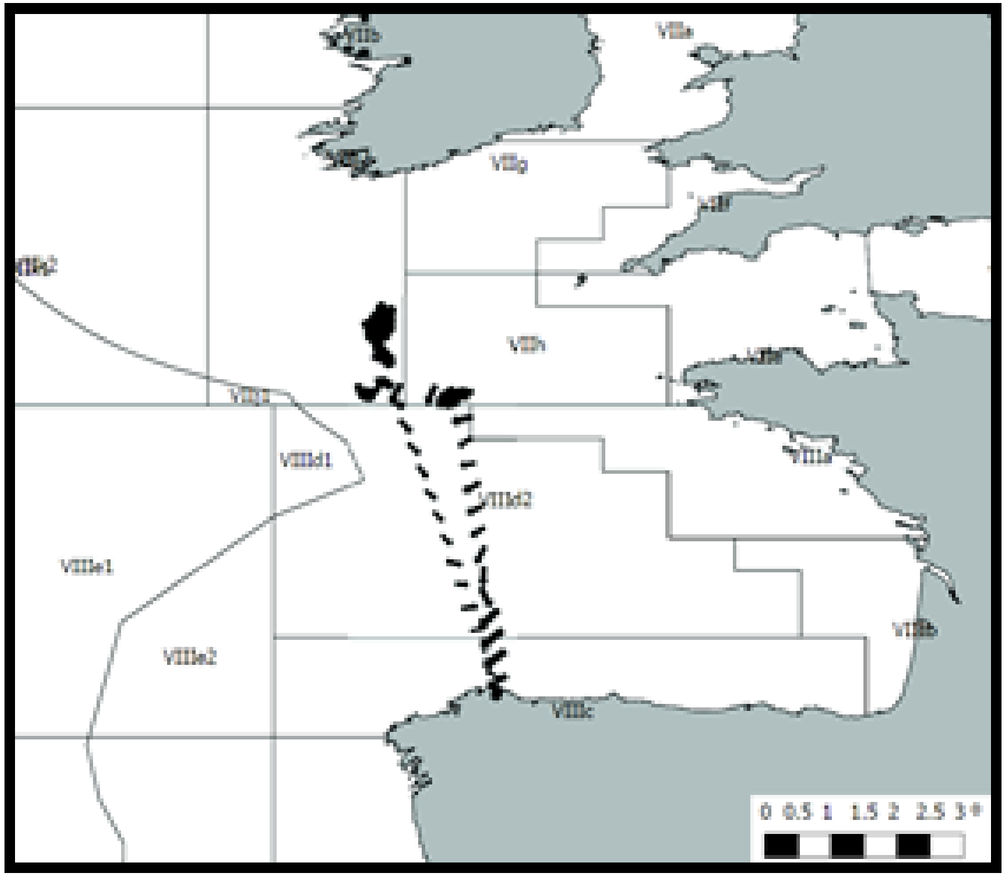


CiTIUS

Datos Espacio-temporales

Objetos Móviles

- Cambio continuo de propiedades geométricas de entidades espaciales
- Tipos
  - Moving Points (muy comunes)
  - Moving Lines
  - Moving Regions
- Representación finita
  - Muestreo de la geometría con una determinada frecuencia temporal
  - Interpolación para instantes intermedios



CiTIUS

Guion

- Introducción
- Datos espacio-temporales
- **Captura de datos**
  - **Procesos de observación (OGC – SensorML)**
  - **Modelo de datos (OGC – O&M)**
- Análisis de datos espaciales
- Tecnologías
- Conclusiones

CiTIUS

Captura de datos

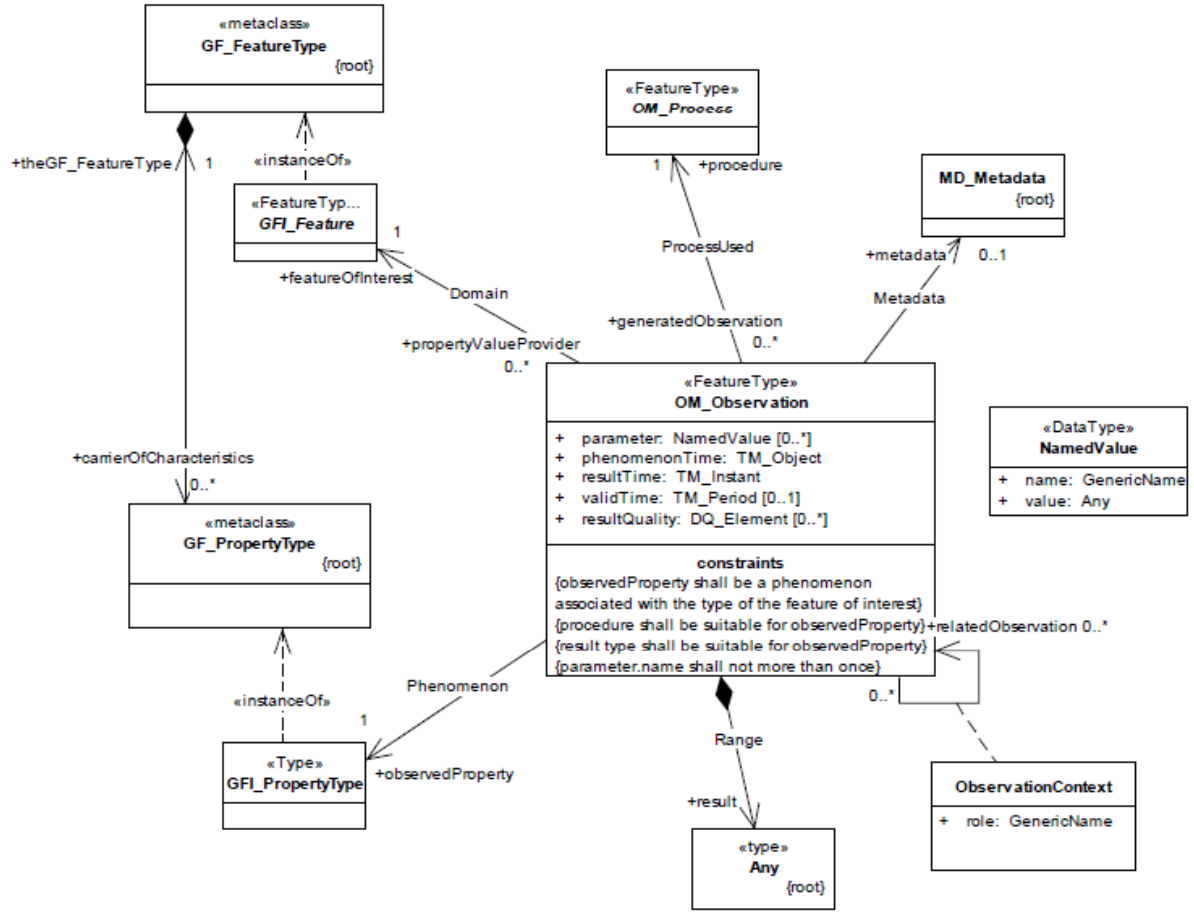
Procesos de Observación

OBSERVACIÓN

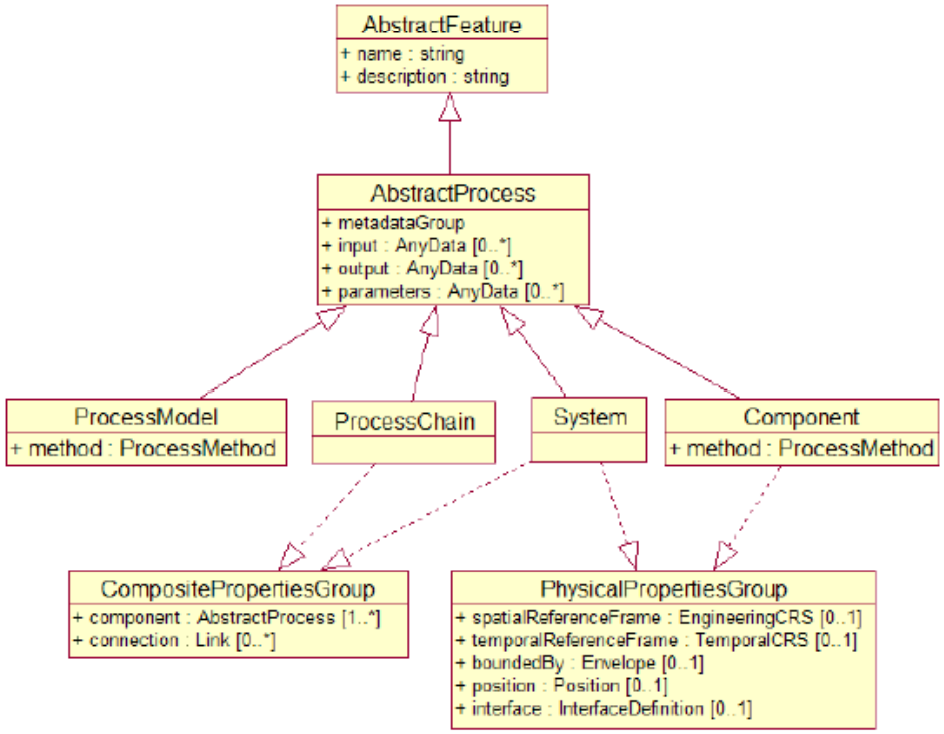
- Proceso (Normalmente Sensor)
- Entidad observada
- Propiedad observada
- Instante temporal
- Valor observado

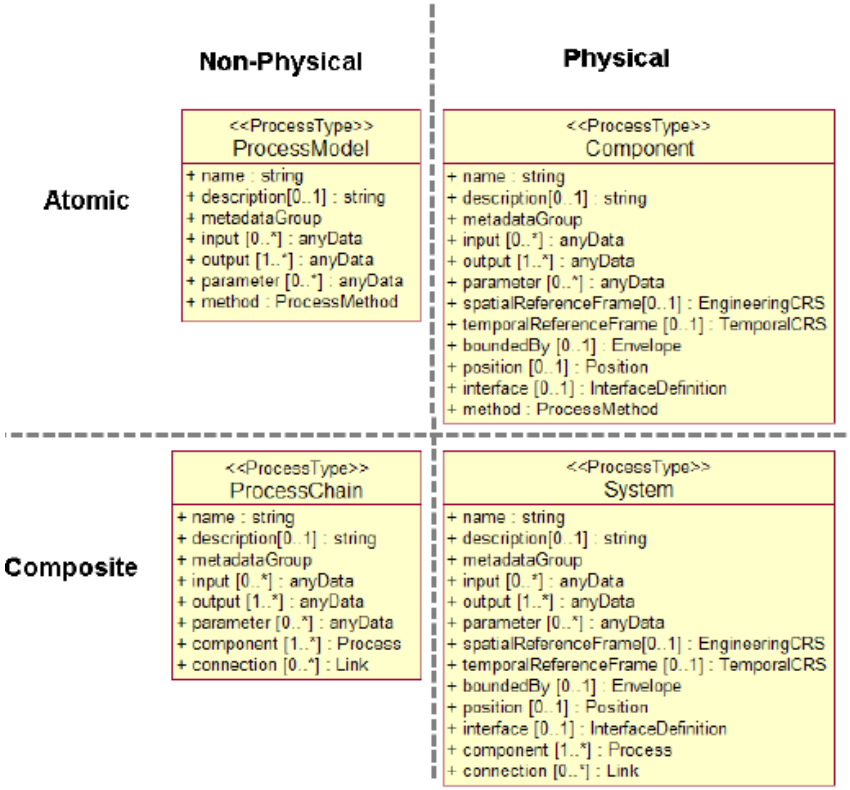
CiTIUS

OGC - OBSERVATIONS & MEASUREMENTS (O&M)



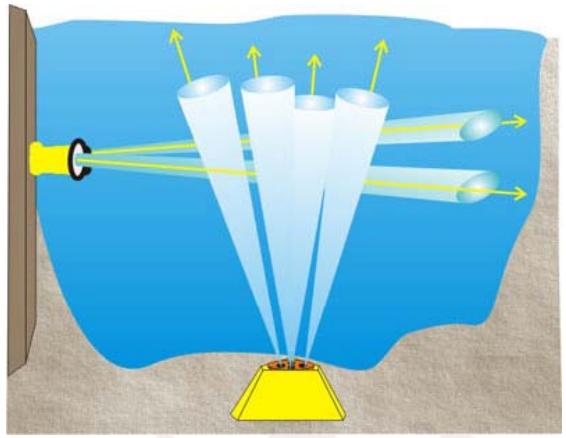
OGC - SENSOR MODEL LANGUAGE (SENSORML)

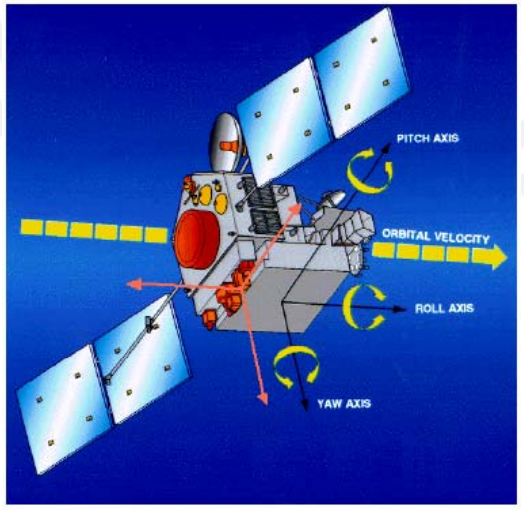




OGC - SENSOR MODEL LANGUAGE (SENSORML)

- **Tipos de Sensores**
  - Sensores In-Situ
  - Sensores Remotos
- **Plataformas**
  - Plataformas estáticas
  - Plataformas móviles





Guion

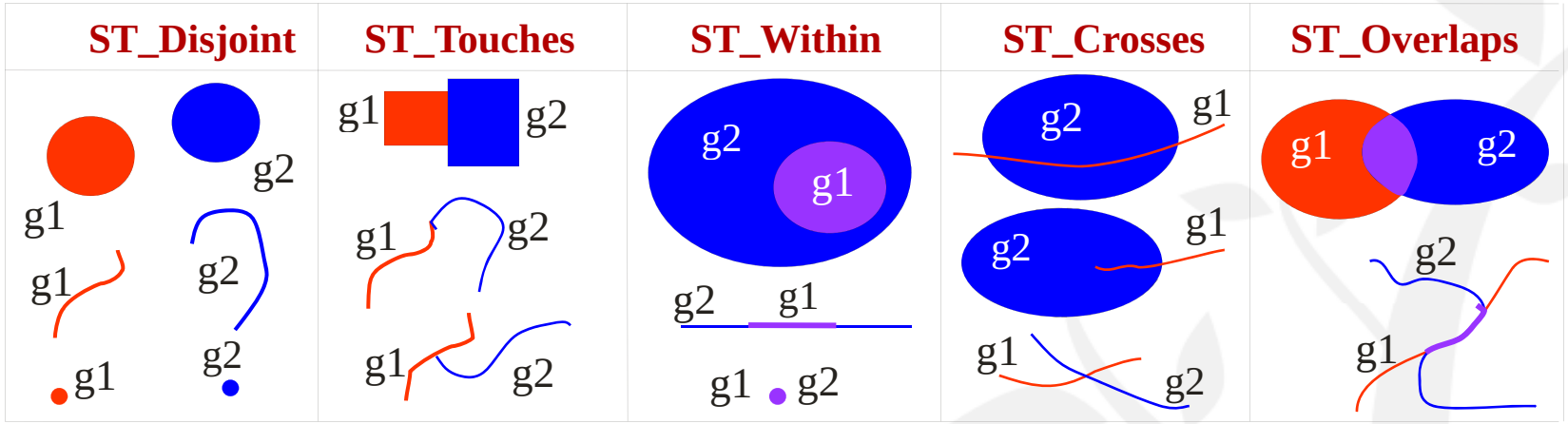
- Introducción
- Datos espacio-temporales
- Captura de datos
- **Análisis de datos espaciales**
  - **Colecciones de entidades espaciales**
  - **Coberturas espaciales**
- Tecnologías
- Conclusiones

CiTIUS

Análisis de datos espaciales

Colecciones de entidades Espaciales

- Predicados topológicos
  - Calculus Based Method (CBM)

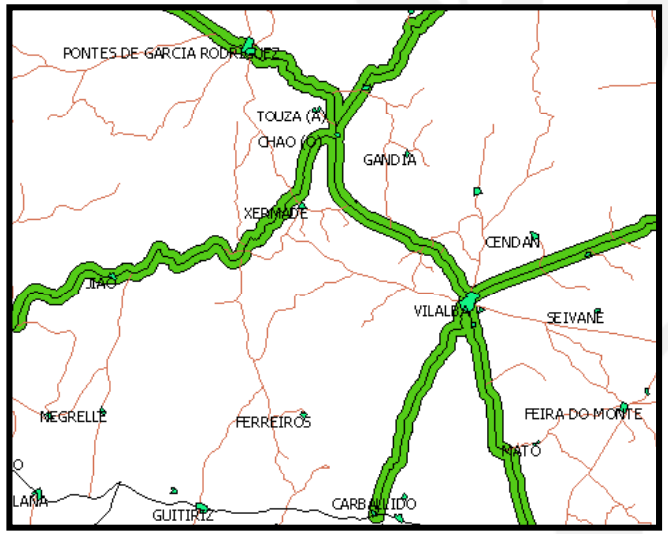


Análisis de datos espaciales

Colecciones de entidades Espaciales

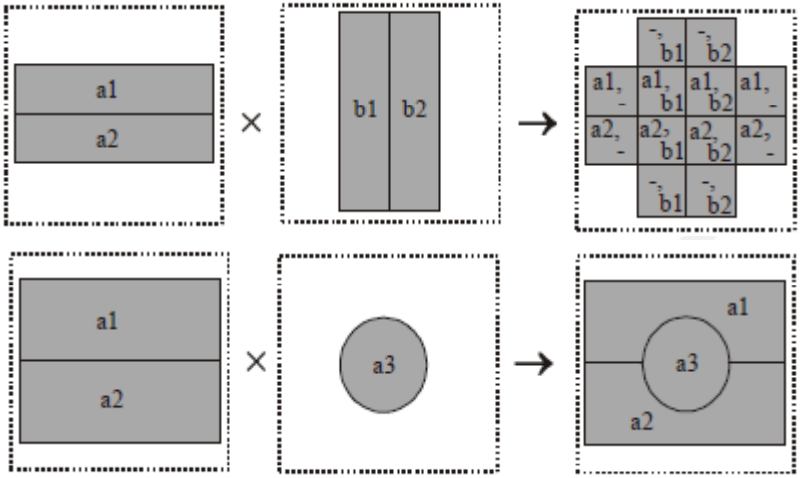
- Propiedades de las geometrías
  - Tipo convencional: area, perímetro, longitud, etc.
  - Tipo espacial: mbr, borde, centroide, etc.
- Cálculo de distancias entre geometrías
- Operaciones de conjunto
  - Unión
  - Intersección
  - Diferencia
- Análisis de redes espaciales
  - Conectividad, caminos más cortos, etc.

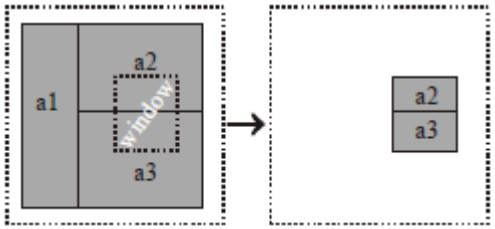
- Rasterización
- Buffer
- Extrusión
- Convex Hull
- Voronoi
- Triangulación de Delaunay
- Etc.

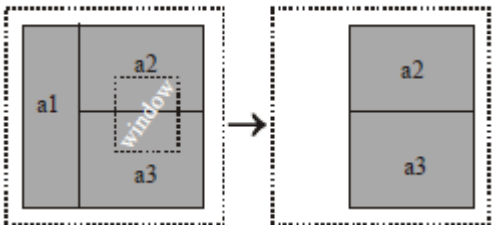


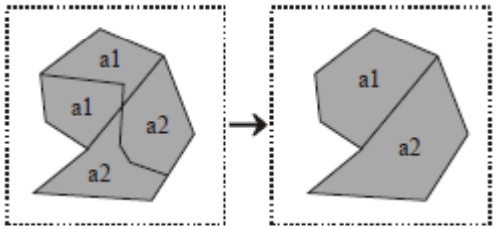
COBERTURAS DISCRETAS (VECTORIAL)

- Gestión de la topología
- Overlay
- Superimposición
- Clipping, Windowing
- Disolución de regiones
- Selección
- Cobertura



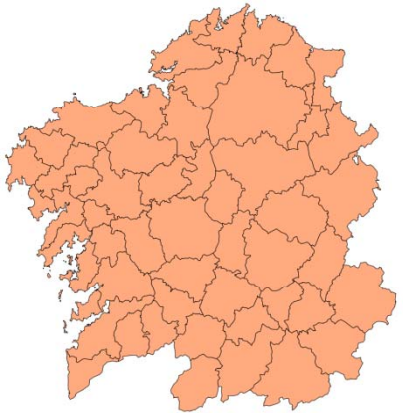


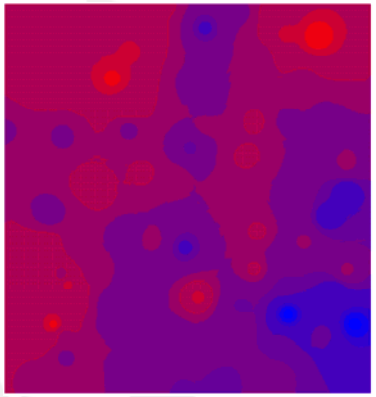


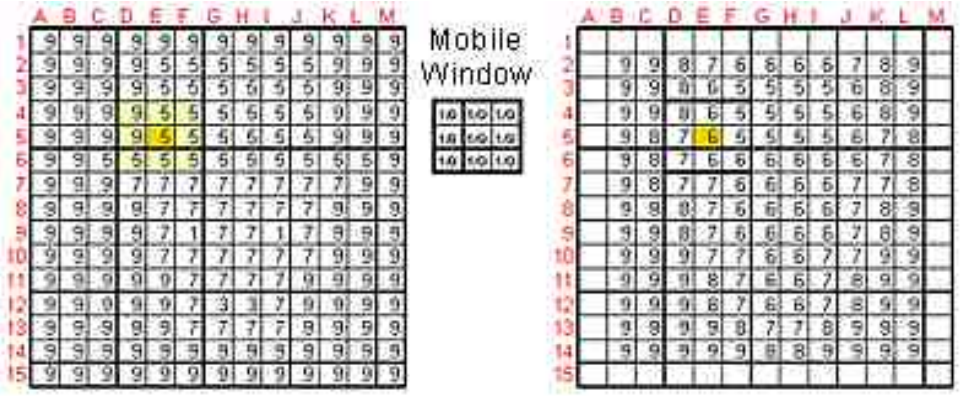


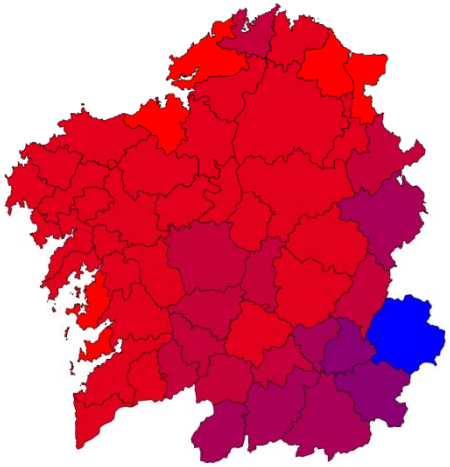
COBERTURAS CONTINUAS (RASTER)

- Algebras de Mapas
  - Basadas en la aproximación de Dana Tomlin
  - Operaciones Locales
  - Operaciones Zonales
  - Operaciones Incrementales
  - Operaciones Focales







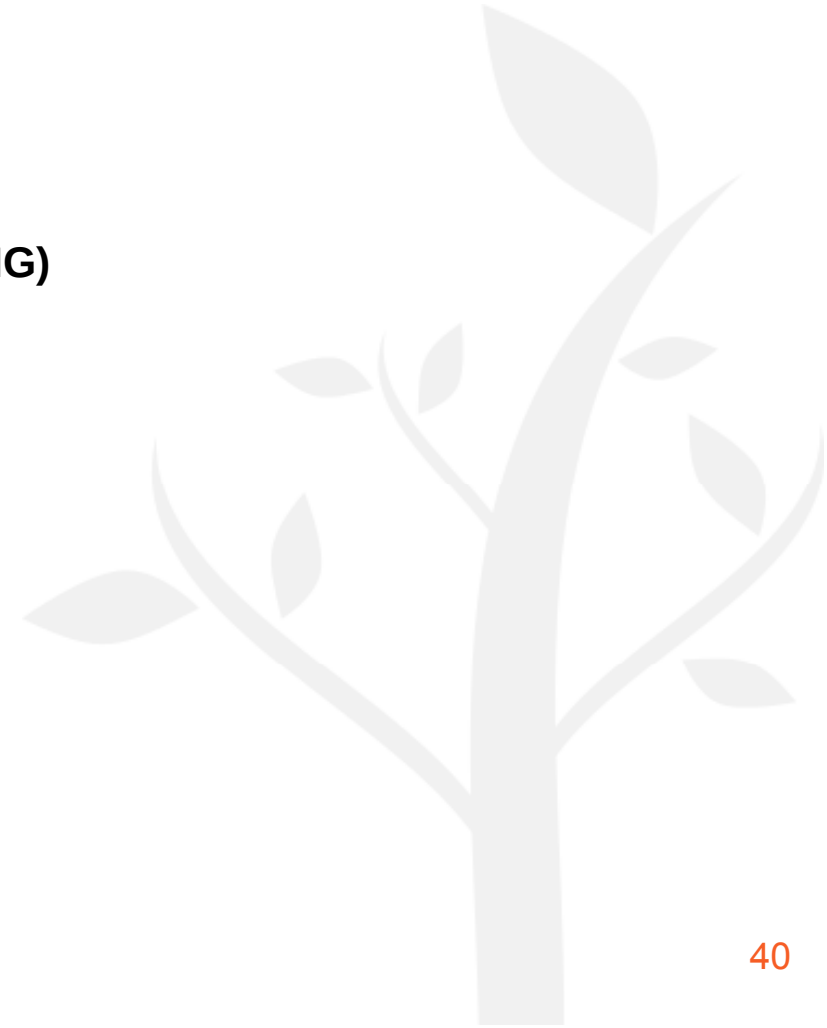


Coberturas Continuas (Raster)

- Generalización
- Vectorización
- Análisis de terrenos
- Modelos Hidrológicos
- Generación de curvas de nivel
- Interpolaciones espaciales
- Transformadas de Fourier
- Álgebra lineal de matrices
- Funcionalidad compleja
  - Modelos meteorológicos
  - Modelos oceanográficos
- Etc.

Guion

- Introducción
- Datos espacio-temporales
- Captura de datos
- Análisis de datos espaciales
- **Tecnologías**
  - **Sistemas de Información Geográfica (SIG)**
  - **Bases de datos espaciales**
  - **Sistemas de gestión de arrays**
- Conclusiones

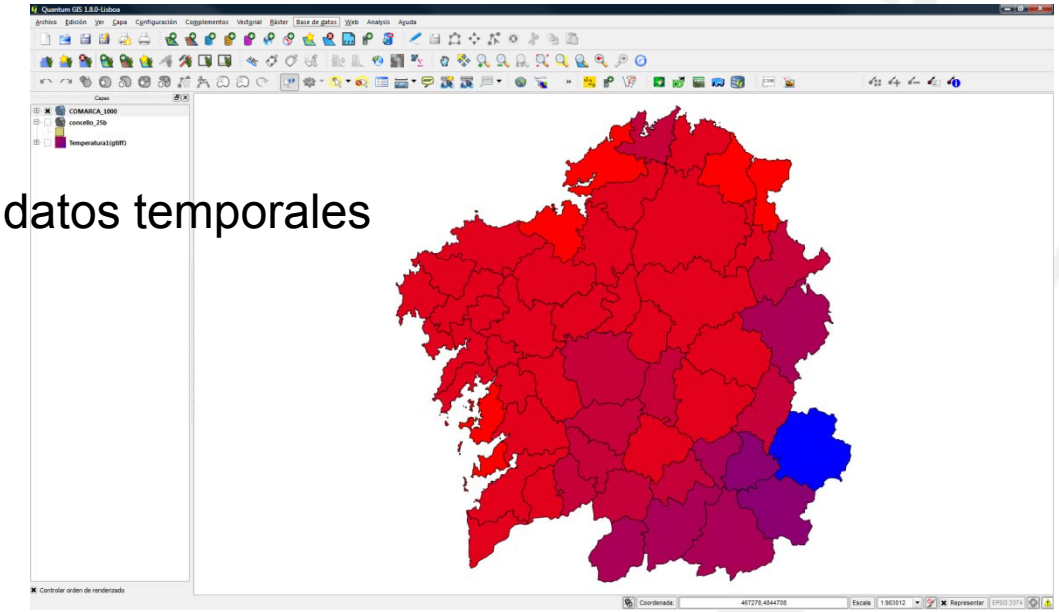


CiTIUS

Tecnologías

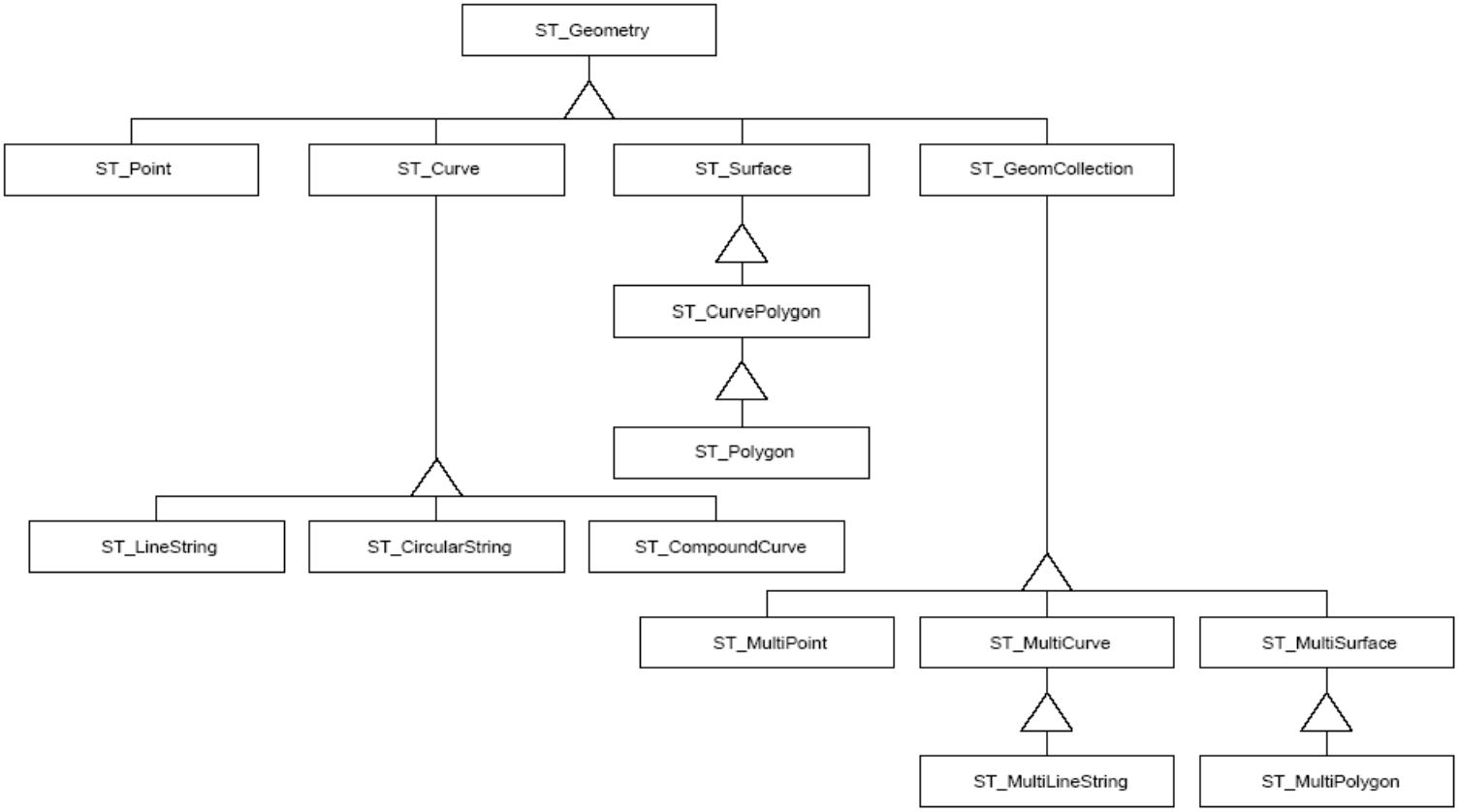
Sistemas de Información Geográfica (SIG)

- Espacio geográfico
- Funcionalidad de análisis espacial de propósito general
- Interfaces de usuario final + entornos de programación
- **No diseñadas para grandes volúmenes de datos**
- Integran fuentes de datos
  - Archivos
  - Bases de datos espaciales
  - Servicios web
- Muy limitadas en la gestión de datos temporales
- Ejemplos
  - gvSIG
  - Quantum GIS
  - ArcGIS
  - Geomedia
  - Etc.



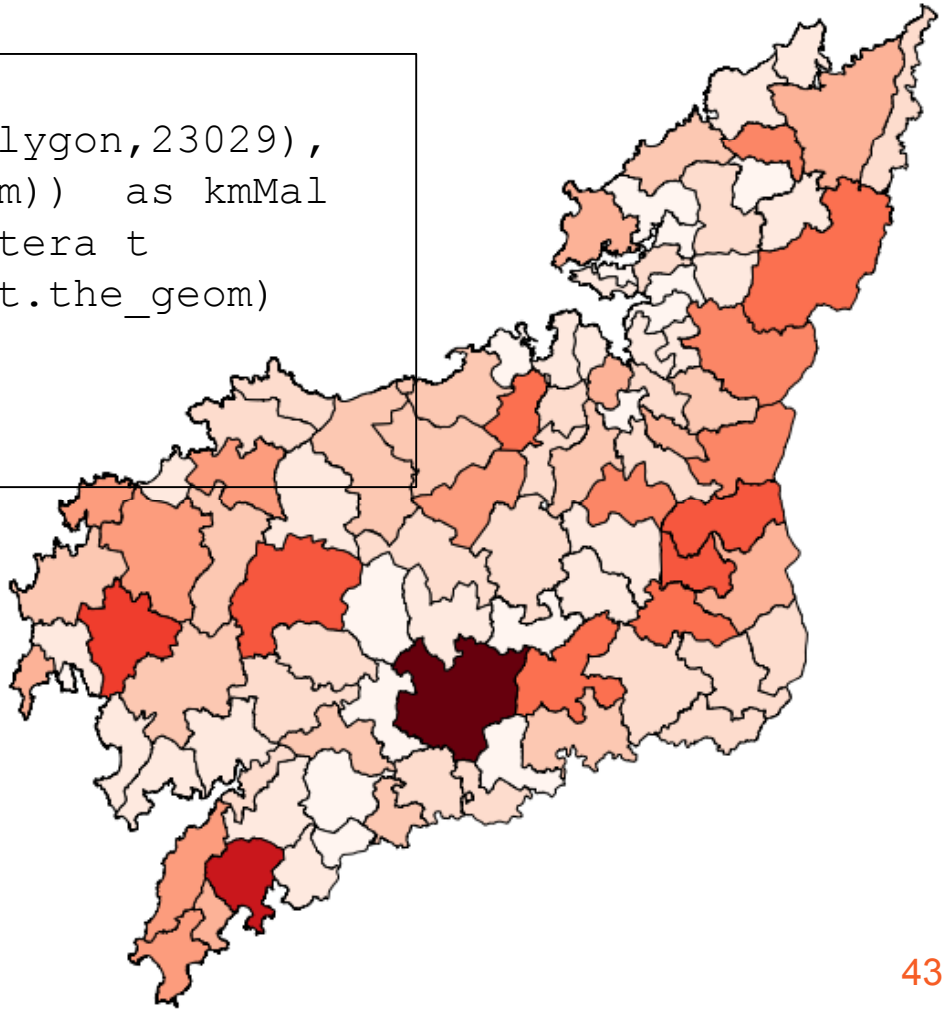
Tipos de Dato Espaciales

Tipos de datos geométricos para colecciones de entidades espaciales
SQL Multimedia and Application Packages (SQL/MM) — Part 3: Spatial



Tipos de Dato Espaciales

```
SELECT m.gid,
       m.the_geom::Geometry(Polygon,23029),
       SUM(st_length(t.the_geom))  as kmMal
FROM municipio m, tramos_carretera t
WHERE st_contains(m.the_geom, t.the_geom)
  AND t.estado = 'M'
GROUP BY by m.gid, m.the_geom
```



Tipos de Dato Temporales

- SQL:2003, SQL:2008
  - TimeStamp, con y sin zona horaria
  - Time, con y sin zona horaria
  - Date
  - Intervalos
    - Año, mes. Ejemplo: 3 años y 10 meses
    - Día, hora. Ejemplo: 123 días, 12 horas, 15 minutos y 5.45 segundos
- SQL:2011
  - Incluye períodos de tiempo de validez y tiempo de transacción a nivel de tupla, que forman parte de la clave (Bitemporal)
  - Predicados temporales para períodos
    - Contains, Overlaps, Precedes, etc.
  - Consulta en un instante específico o en un período de tiempo de transacción

IMPLEMENTACIONES

- SGBD de almacenamiento por filas
  - Implementan estándares de ISO
    - Parte espacial del SQL/MM
    - Pronto parte temporal del SQL:2011
  - Oracle Spatial (http://www.oracle.com/es/products/database/options/spatial/index.html)
  - IBM DB2 Spatial Extender (http://www-03.ibm.com/software/products/us/en/db2spaext/)
  - PostgreSQL + PostGIS (http://www.postgis.org/)
  - Algunos incluyen tipos de datos para almacenar Raster
    - Oracle, PostGIS
- SGBD de almacenamiento por columnas
  - Extensión geospacial de MonetDB (http://www.monetdb.org/Documentation/Extensions/GIS)
  - Implementa la especificación OGC – Simple Features for SQL

IMPLEMENTACIONES

- Bases de datos NoSQL
  - Implementaciones todavía muy limitadas en funcionalidad
  - Gestión de documentos GeoJSON (http://geojson.org/)
    - Formato para codificar en JSON colecciones de entidades espaciales con propiedades de tipos geométricos
    - MongoDB (http://www.mongodb.org/)
    - Geocouch (https://github.com/couchbase/geocouch/)
  - Seguramente pronto veremos más
    - Cassandra
    - voltDB (new SQL)

Tecnologías

Sistemas de gestión de arrays

ARCHIVOS

- Formatos para imagen geográfica
  - GeoTIFF (formato TIFF con cabecera geográfica)
  - ecw (Enhanced Compressed Wavelet)
- Arrays e imágenes en entornos científicos (Ejemplos)
  - Flexible Image Transport System (FITS)
  - Hierarchical Data Format (HDF4, HDF5)
  - GRIdded Binary or General Regularly-distributed Information in Binary form (GRIB)
  - Network Common Data Form (NetCDF)
- BIG Data?
  - Hadoop + HDFS
    - Implementaciones ad-hoc

CiTIUS

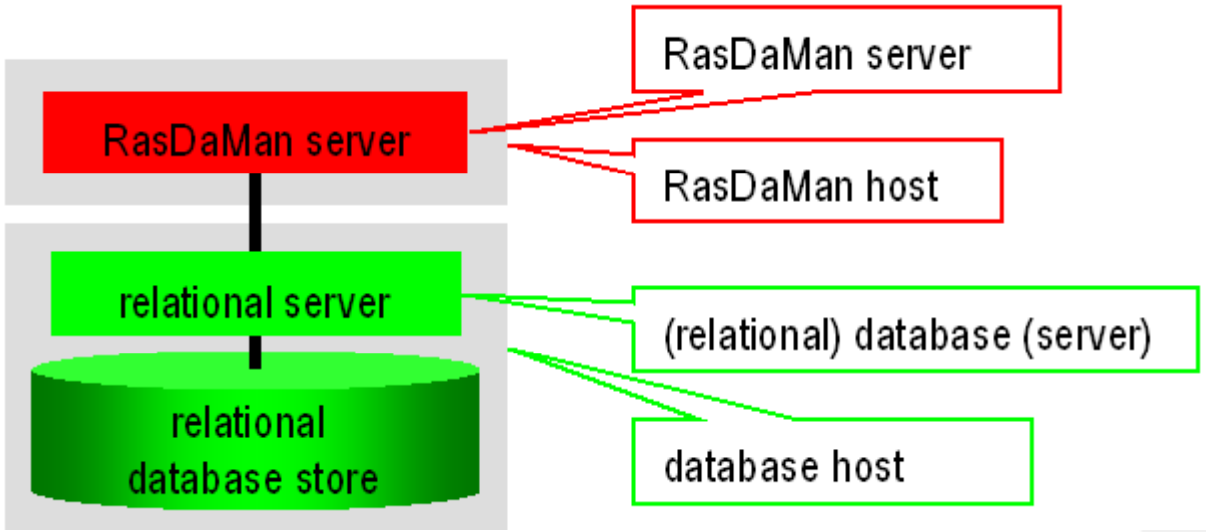
BASES DE DATOS OBJETO-RELACIONALES (SQL:2003)

- Tipos de dato ARRAY
- Permiten anidar arrays dentro de tuplas del modelo relacional
- No diseñado para gestionar los arrays de gran tamaño que aparecen en aplicaciones de tipo científico

```
CREATE TABLE Empleado (
  id_emp INTEGER PRIMARY KEY,
  nombre VARCHAR(50),
  salarios DECIMAL(8,2) ARRAY[12]
)
```

RASDAMAN (http://www.rasdaman.org/)

- Gestor de arrays multidimensionales
- Implementado sobre un gestor relacional
  - Típicamente PostgreSQL
  - Arrays almacenados en tipos BLOB



RASDAMAN (http://www.rasdaman.org/)

- Lenguaje de consulta declarativo sobre arrays (**rasql**)
  - Lenguaje de definición de arrays (**rasdl**)
    - Definición del tipo de las celdas del array
      - Tipos simples y Tipos complejos (struct)
    - Definición de arrays sobre un tipo de celda
    - Definición de colecciones de arrays del mismo tipo

```
struct RGBPixel {char red, green, blue; };

typedef marray <RGBPixel,[0:799, 0:599]> RGBImage;

typedef set <RGBImage> RGBSet;
```

RASDAMAN (http://www.rasdaman.org/)

- Lenguaje de consulta declarativo sobre arrays (**rasql**)
  - Lenguaje de manipulación de arrays (**rasml**)
  - Similar a SQL.

```
select RGBSet[ 120:160, 55:75 ]
from RGBSet
```

```
select RGBSet.red * 2
from RGBSet
```

```
select RGBSet
from RGBSet
where all_cells(RGBSet.green > 20)
```

```
select png(RGBSet)
from RGBSet
```

```
select marray v in [0:255]
       values condense +
              over x in sdom(RGBSet)
              where RGBSet[x].green = v[0]
              using 1
from RGBSet
```

CiTIUS

RASDAMAN (http://www.rasdaman.org/)

- Lenguaje de consulta declarativo sobre arrays (**rasql**)
  - Lenguaje de manipulación de arrays (**rasml**)
  - Similar a SQL.

```
select RGBSet[ 120:160, 55:75 ]
from RGBSet
```

```
select RGBSet.red * 2
from RGBSet
```

```
select RGBSet
from RGBSet
where all_cells(RGBSet.green > 20)
```

```
select png(RGBSet)
from RGBSet
```

```
select marray v in [0:255]
       values condense +
              over x in sdom(RGBSet)
              where RGBSet[x].green = v[0]
              using 1
from RGBSet
```

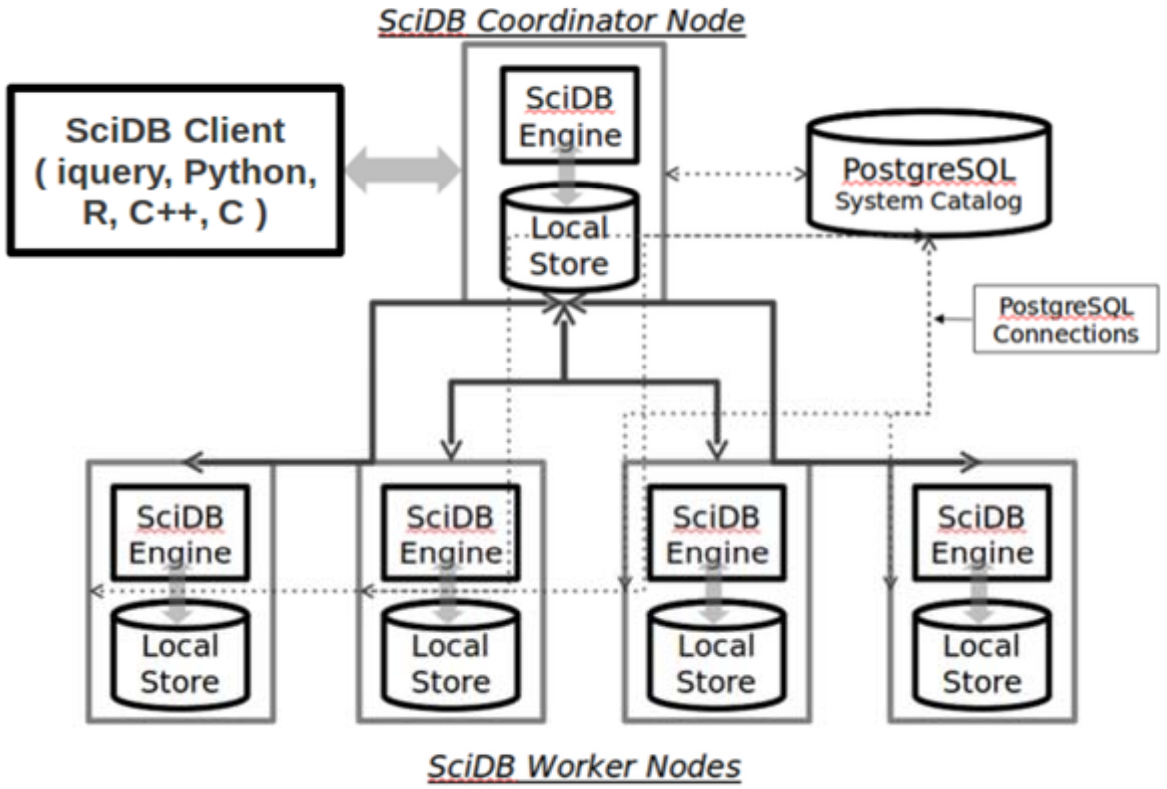
CiTIUS

SciDB(http://www.scidb.org/)

- Optimizado para Big Data y Big Analytics
- Arrays y Vectores son objetos de primera clase con operadores específicamente optimizadas para ellos.
- Arquitectura paralela de tipo share-nothing
- Extensible: Tipos y funciones definidos por el usuario
- Interacción con el paquete estadístico R
- Multiversión (tiempo de transacción)
- Comparativa entre Hadoop + HDFS y SciDB realizada por la NASA
  - Respective Strengths and Weaknesses of SciDB, MapReduce-HDFS, and a Custom Technique for a Data-Intensive Analysis System
  - https://ams.confex.com/ams/93Annual/flvgateway.cgi/id/23219?recordingid=23219
  - Cálculo de agregados sobre 6GB de datos
    - Hadoop: 180 segundos. 89 segundos después de optimizar
    - SciDB: 39 segundos out-of-box

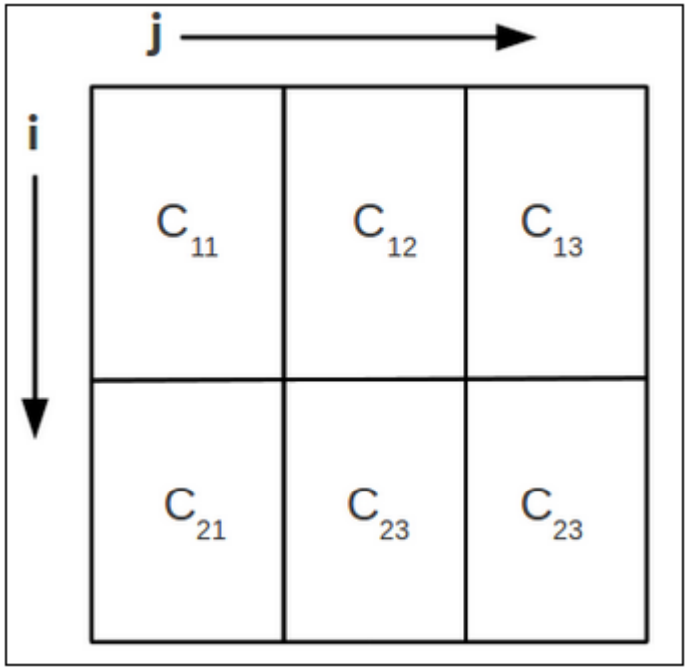
SciDB(http://www.scidb.org/)

- Arquitectura



SciDB(http://www.scidb.org/)

- Modelo de datos de arrays
  - Array
    - Nombre
    - Dimensiones
    - Atributos
- Chunking
  - Cada dimensión se divide en *Chunks*
    - Dimensión i con tamaño 10 y tamaño de chunk 5
    - Dimensión j con tamaño 30 y tamaño de chunk 10
  - Los chunks se distribuyen por los nodos usando un esquema basado en hashing
  - Solapamiento de los chuncks
    - `A <a: int32> [i=1:10,5,1, j=1:30,10,5]`
    - Mejora rendimiento en operaciones de vecindario



SciDB(http://www.scidb.org/)

- Almacenamiento
  - Dentro de cada chunk se almacena por atributos (particionamiento vertical)
  - Compresion con Run Length Encoding (RLE) en cada atributo
  - Los datos nunca se sobrescriben, se generan nuevas versiones
- Catalogo almacenado en PostgreSQL
- Lenguajes
  - Array Query Language (AQL)
    - DDL + DML
  - Array Functional Language (AFL)
    - Operadores sobre arrays que incluyen Algebra Lineal de vectores y matrices

```
CREATE ARRAY A
   <val_a:double>[i=0:9,10,0];

SELECT sqrt(val_b)
FROM A
WHERE j>3 AND j<7
```

				4.3	3.4	3.1			

Guion

- Introducción
- Datos espacio-temporales
- Captura de datos
- Análisis de datos espaciales
- Tecnologías
- **Conclusiones**

CiTIUS

Conclusiones

- Gestión de datos espaciales y espacio-temporales presente en un gran número de dominios de aplicación
  - Muchas aplicaciones de gestión de datos científicos
  - Se necesitan extensiones espaciales
- Gestión de entidades espaciales resuelto con bases de datos espaciales
  - Generalmente no necesitan tecnologías de Big Data
- Big Data?
  - Almacenamiento y análisis de datos de sensores
- Herramientas SIG no soportan grandes cantidades de datos
- Almacenamiento de datos crudos de sensores
  - Muchas inserciones por unidad de tiempo
    - Aproximaciones NoSQL: Cassandra, etc.
    - Nuevas soluciones SQL: voltDB, etc.
- Análisis sobre muestreos temporales y espaciales (coberturas)
  - Tecnologías de gestión de arrays multidimensionales

CiTUS

Gracias por su atención

José Ramón Ríos Viqueira: jrr.viqueira@usc.es

citius.usc.es