Algoritmo para la detección automática de entidades de población*

Miguel R. Luaces, Isabel Pérez-Urria Lage, David Trillo Pérez

Laboratorio de Bases de Datos. Universidade da Coruña
Campus de Elviña S/N. 15071 A Coruña, España
{luaces, iperezurria, dtrillo}@udc.es

Resumen

El proyecto *Diagnóstico del Hábitat rural de Galicia* surge de un convenio entre la *Consellería de Vivenda e Solo* de la *Xunta de Galicia* y la Universidad de A Coruña con el objetivo de recuperar y rehabilitar el hábitat rural de Galicia. Para ello es necesario realizar un estudio que permita analizar el estado actual de las entidades de población rurales de Galicia; el análisis detallado de los resultados permitirá revitalizar los asentamientos tradicionales de población y evitar el despoblamiento del medio rural, revalorizando el patrimonio arquitectónico gallego. Para la delimitación de las entidades de población encuestadas en este proyecto se diseñó un algoritmo de detección de entidades de población que identifica agrupaciones de casas que sean susceptibles de ser núcleos rurales, para ello estas agrupaciones deben cumplir ciertos requisitos, como puede ser que el número de edificaciones se encuentre acotada por un mínimo y un máximo, o que las edificaciones no se encuentren demasiado dispersas. En este artículo se describe en primer lugar el problema y las dos alternativas existentes a la hora de abordarlo. A continuación se presenta de forma detallada el algoritmo que hemos diseñado para resolver el problema y su implementación utilizando software libre. Para finalizar, se presenta brevemente el proyecto en el que se enmarca este trabajo.

Palabras clave: SIG, análisis territorial, detección de entidades, software libre.

1 Introducción

* Este trabajo ha sido parcialmente financiado por el "Ministerio de Educación y Ciencia" (PGE y FEDER) ref. TIN2006-16071-C03-03, por la "Agencia Española de Cooperación Internacional (AECI)" ref. A/8065/07, y por la "Xunta de Galicia" ref. 2006/4 y ref. 08SIN009CT.

Uno de los objetivos de la *Consellería de Vivenda e Solo* de la *Xunta de Galicia* es la recuperación y rehabilitación del hábitat rural de Galicia. Para ayudar en la consecución de este objetivo la *Consellería* firmó un convenio con la Universidad de A Coruña para realizar un estudio que permita analizar el estado actual de las entidades de población rurales de Galicia. Este proyecto se denomina *Diagnóstico del Hábitat rural de Galicia*, y en él participan investigadores de la Escuela Técnica Superior de Arquitectura y del Laboratorio de Bases de Datos de la Facultad de Informática. El análisis detallado de los resultados permitirá revitalizar los asentamientos tradicionales de población y evitar el despoblamiento del medio rural, revalorizando el patrimonio arquitectónico gallego.

La primera etapa del proyecto consiste en abordar 12 comarcas gallegas que suman una superficie aproximada de 7.286 km2 y unos 375.000 habitantes. Una de las tareas críticas de esta primera etapa consiste en determinar las entidades de población que posteriormente serán encuestadas por los equipos de recogida de datos.

En una primera aproximación, y siguiendo el ejemplo de Asturias, comunidad en la que se llevó a cabo un proyecto similar, se decidió abordar la tarea de detección de entidades de forma manual. El método consistía en la división del territorio en cuadrículas UTM y la delimitación de entidades por inspección visual, en base a parámetros como número de edificaciones o distancia entre las mismas. Sin embargo, pronto surgió la certeza de que esta tarea sería inabordable debido la gran superficie a tratar y el carácter disperso del hábitat gallego. Además, este método no asegura la homogeneidad y la coherencia entre distintas zonas porque el resultado es completamente subjetivo. La segunda aproximación que se siguió para la resolución del problema de delimitación de entidades fue la utilización de un sistema de información geográfica (GIS) y el diseño de un algoritmo automático que determinarse las entidades que debían ser encuestadas. Para ello, existen dos alternativas: utilizar un algoritmo orientado a cuadrículas o utilizar un algoritmo de detección de entidades de población.

En los algoritmos orientados a cuadrículas la región geográfica a encuestar se divide en una cuadrícula y se determina si cada celda de la cuadrícula debe ser encuestada o no en función de ciertos parámetros geográficos de la información contenida en la celda (número de edificaciones, densidad de población, etc.). Las ventajas de emplear cuadrículas para la recopilación de datos estadísticos de regiones geográficas se destacan en estudios realizados en Noruega, Finlandia y Suecia [1]. Estos estudios afirman que una ventaja de las cuadrículas es que se

mantienen espacialmente estables, a diferencia de las entidades delimitadas automáticamente que son dependientes de los cambios regionales. Además, los datos de cuadrículas pequeñas pueden ser sumados para obtener áreas más grandes. Sin embargo, tienen el problema de que la cuadrícula pueden dividir una zona de interés en dos celdas y considerarla independiente. Podría ocurrir, por ejemplo, que una entidad de población que si estuviera contenida en una celda sería encuestada, quede separada en dos celdas y no sea encuestada.

Por otra parte, los algoritmos de detección de entidades de población determinan los límites de las entidades de población que deben ser encuestadas a partir de la propia información geográfica (edificaciones, viario, etc.) mediante un algoritmo constructivo que se aplica sobre toda la superficie a encuestar. Estos algoritmos utilizan los objetos geográficos (polígonos, líneas y puntos), sus relaciones topológicas (p.e., solapamiento) y métricas (p.e., distancia o superficie), y la información alfanumérica asociada a los objetos para determinar la ubicación geográfica de las entidades de población que deben ser encuestadas. Los criterios empleados, entre otros, son el número de habitantes, la distancia entre edificaciones o que las edificaciones estén bien conectadas por carreteras. Podemos encontrar otros ejemplos en los que la densidad de población es un parámetro importante en el método de delimitación, como el del Instituto de Medio Ambiente y Sostenibilidad de Ispra (Italia), en el que se intenta realizar una clasificación de los núcleos en urbanos y rurales en base a la densidad de población, la proporción de tierra de cultivo o áreas artificiales [2]. La gran ventaja de estos métodos es que no tienen el problema de las cuadrículas. Sin embargo, su gran problema es que son más complejos de diseñar, implementar y ejecutar.

Comparando estos casos de estudio con el nuestro encontramos varias diferencias. En primer lugar, la asociación de información a cuadrículas tiene menos utilidad práctica que asociarla directamente a entidades, sobre todo a la hora de visualizarla, realizar búsquedas o análisis estadísticos. Por ello, desechamos rápidamente la utilización de un algoritmo orientado a cuadrículas. En segundo lugar, en ambos métodos se tienen en cuenta parámetros como el número de habitantes, o la densidad de población. En nuestro caso esto no tiene sentido pues deseamos detectar igualmente los núcleos abandonados o semi-abandonados que tienen una densidad de población próxima a cero. Por ello fue necesario diseñar un nuevo algoritmo que tuviera en cuenta las características especiales de nuestro problema.

El resto de este artículo se estructura de la siguiente forma. En primer lugar se presenta de forma detallada el algoritmo que hemos diseñado para resolver el problema y su implementación utilizando software libre. A continuación, se

describe de forma breve la forma en la que este algoritmo se usa en la metodología general del proyecto.

2 Algoritmo de detección de entidades del proyecto Hábitat

Para la delimitación de las entidades de población encuestadas en este proyecto se diseñó un algoritmo de detección de entidades de población que emplea como información geográfica de partida la capa vectorial de edificaciones obtenida de la cartografía del SITGA (Sistema de Información Territorial de Galicia, dependiente de la Xunta de Galicia). El algoritmo detecta agrupaciones de casas que sean susceptibles de ser núcleos rurales. Para ello, estas agrupaciones deben cumplir ciertos requisitos como puede ser que el número de edificaciones se encuentre acotada por un mínimo y un máximo, o que las edificaciones no se encuentren demasiado dispersas.

En base a estos criterios identificamos los siguientes parámetros de entrada del algoritmo:

1. *Distancia máxima entre edificaciones*. Representa la distancia máxima en metros que puede existir entre las edificaciones para considerar ese agrupamiento de viviendas una entidad a visitar.
2. *Número mínimo de edificaciones*. Representa el número mínimo de edificaciones en la entidad para que se considere que esta tiene el tamaño adecuado para ser visitada.
3. *Número máximo de edificaciones*. Este parámetro sirve para descartar aquellas entidades demasiado grandes que tendrían por tanto ya una consideración de hábitats urbanos y no rurales , por lo que quedarían fuera del ámbito de este estudio.

Dado que el objetivo del estudio no se refiere únicamente al interior de las entidades, sino que el entorno de las mismas también es determinante para clasificarlo, en un último paso se procederá a calcular una zona de influencia alrededor de la superficie de cada entidad. Además, para la asociación posterior de topónimos a las entidades también se utilizó la capa vectorial de nombres de lugar.

0 **1**

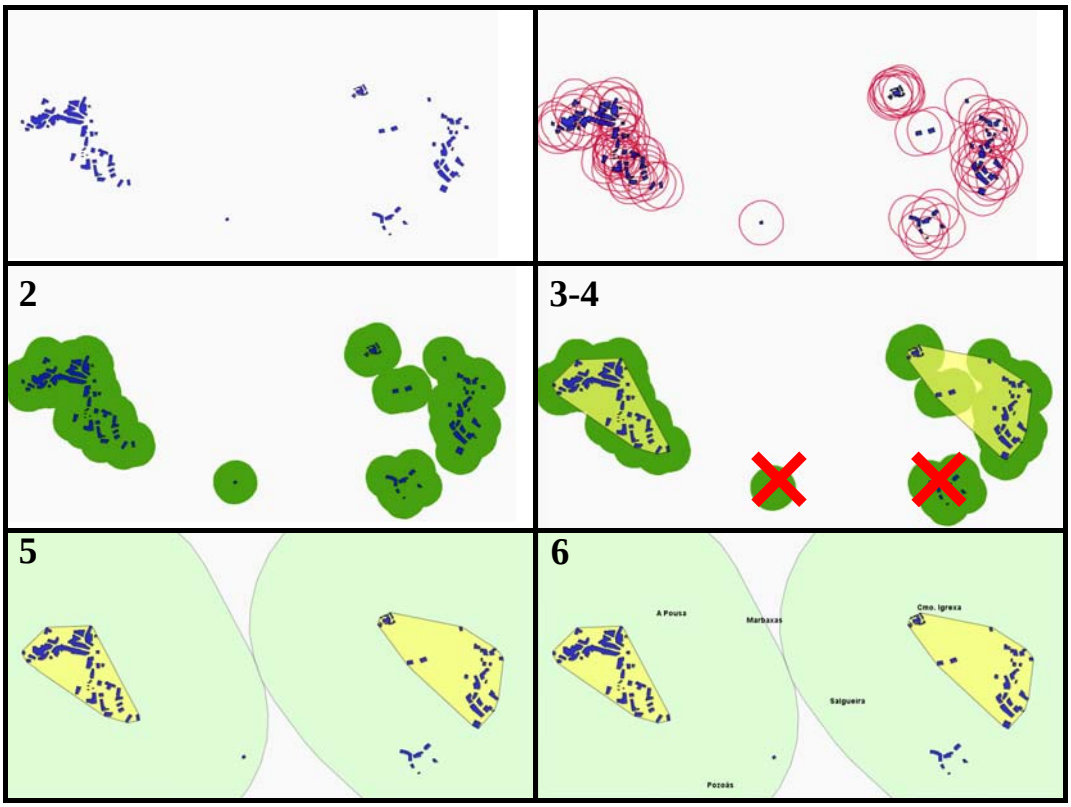

Figura 1: Ejemplo de funcionamiento del algoritmo de detección de entidades

A partir de estos parámetros se diseñó un algoritmo que hace uso de operaciones espaciales para la detección de entidades. En la Figura 1 se muestra el funcionamiento del algoritmo de detección de entidades, sobre unas agrupaciones de casas, con los siguientes parámetros de entrada:

- Distancia máxima de edificaciones: 100 m
- Número mínimo de edificaciones: 10
- Número máximo de edificaciones: 100
- Distancia máxima de la zona de influencia: 300 m

El algoritmo consta de los siguientes pasos:

1. Creación de un *buffer* alrededor de todas las edificaciones cuyo ancho es la distancia máxima entre edificaciones dividida por dos.

2. Disolución los *buffers* de edificaciones (realizar la unión de aquellos que intersecan).
3. Eliminación de aquellos *buffers* disueltos cuyo número de edificaciones no se encuentre entre los límites especificados.
4. Cálculo del *convex-hull* de las edificaciones contenidas en los *buffers* disueltos.
5. Cálculo de la zona de influencia de cada entidad. Para ello se creará un *buffer* alrededor de la superficie calculada para cada entidad.
6. Asociación de un topónimo a cada entidad detectada:
    a. Para cada entidad se obtuvieron varias denominaciones candidatas, escogidas por su proximidad a la entidad.
    b. Las denominaciones se ordenan ascendentemente por distancia a la entidad
    c. Los equipos de recogida de datos determinarán la denominación correcta.

Hay que precisar que en la cartografía del SITGA un polígono de la capa de edificaciones puede representar una o varias edificaciones en la realidad, por lo que se consideró que los polígonos que superaban un umbral representaban varias edificaciones. En concreto, se aplicó la siguiente estimación: Si un bloque tiene hasta 250 m^2 de superficie, es considerado una única edificación, a partir de esa superficie, por cada 200 m^2 de bloque sumamos una edificación más. De esta forma, un bloque de 500 m2 será considerado como 3 edificaciones, uno de 1300 m^2 como 7, etc.

Los parámetros de entrada al algoritmo sólo tienen en cuenta valores cuantitativos (distancias, número de edificaciones), por este motivo el algoritmo detecta algunas entidades no relevantes para el estudio (por ejemplo, polígonos industriales, urbanizaciones, etc.). Es preciso realizar un filtrado manual a posteriori para eliminar todo aquello que no sean núcleos puramente rurales. La utilización durante el desarrollo del proyecto de una aplicación SIG que disponía de un cliente WMS (Web Map Service) [3] permitió la generación de un mapa por entidad con una ortofoto obtenida del WMS del PNOA (Plan Nacional de Ortografía Aérea) que facilitó considerablemente este proceso de filtrado.

Por otra parte, ha de puntualizarse que este método automático de delimitación de entidades no coincide con la delimitación tradicional de núcleos de población y las áreas resultantes constituyen únicamente una abstracción que permite englobar una agrupación de edificaciones, de forma que se puedan recoger datos sobre las mismas y asociarlos directamente a la geometría obtenida.

Para la obtención de las entidades finales se realizaron varios análisis y comparativas empleando distintos valores de los parámetros de entrada. El criterio de selección de parámetros se basó en la observación visual de las entidades obtenidas y que el número de entidades obtenido se aproximase lo más posible al número estimado.

3 Implementación del algoritmo de detección de entidades

El algoritmo de detección de entidades fue desarrollado como una extensión de *gvSIG 1.0* (aplicación de software libre desarrollada para la Comunidad Valenciana) utilizando tecnología *Java* y el sistema gestor de bases de datos *PostgreSQL 8.1* junto con su extensión espacial *Postgis*.

La implementación, realizada totalmente con software libre, utiliza funciones *Postgis* para manejar las geometrías de las entidades implicadas en los cálculos espaciales del algoritmo. Los operadores espaciales *buffer*, y *geomunion* permiten crear los *buffers* sobre las edificaciones y disolverlos. El predicado espacial *not disjoint* es utilizado para calcular el número de edificaciones contenidas en cada buffer disuelto. Y, una vez descartados aquellos buffers cuyo número de edificaciones no está entre los límites, el operador *convexhull* permite delimitar todas las agrupaciones de edificaciones que son consideradas una entidad de población. Finalmente, el operador *buffer* es utilizado de nuevo para calcular la zona de influencia de cada entidad. Además las funciones *collect y multi* son utilizadas para realizar transformaciones en los tipos de datos de las geometrías obtenidas en los cálculos espaciales.

Debido a la complejidad de la operación *geomunión,* el tiempo de disolución de los *buffers* aumentaba considerablemente cuando se procesaban muchas edificaciones en una misma ejecución del algoritmo volviéndose muy ineficiente. Como optimización, la operación de disolución de *buffers* de edificaciones, pasó a realizarse de forma jerárquica comenzando con la disolución de los *buffers* de las edificaciones de un mismo municipio, continuando con la disolución de *buffers* a nivel de comarca y finalmente con la de los *buffers* de todo el espacio de trabajo.

4 Metodología del proyecto

Podemos describir los objetivos concretos del proyecto *Diagnóstico del Hábitat rural de Galicia* como "*la obtención de los datos necesarios que permitan el modelado científico del hábitat de Galicia actual*". Estos datos son referidos a:

1. Identificación, evaluación y localización geográfica de los fenómenos derivados de la construcción del hábitat y de la sensación transmitida al ambiente por la edificación en función de parámetros visuales.
2. Identificación y localización geográfica de las edificaciones de tipología tradicional, evaluando su estado actual.
3. Detección del estado de uso del parque inmobiliario, identificando abandono, ausencia de uso, infravivienda o viviendas inacabadas en cualquiera de sus estados intermedios.
4. Localización de zonas homogéneas en las que desarrollar acciones de segundo orden.
5. Elaboración de conclusiones a partir de los materiales obtenidos, que sinteticen un diagnóstico y que enuncien recomendaciones para las administraciones y, en algunos casos, propuestas de realización de estudios detallados de aspectos o ámbitos singulares.

En la elaboración de las conclusiones, se realizará un análisis detallado de las entidades encuestadas que incluirá un análisis DAFO (Debilidades, Amenazas, Fortalezas y Oportunidades) de cada una de ellas.

La metodología de trabajo se representa en el diagrama de la figura 2. Los pasos de esta metodología son los siguientes:

1. En una primera fase del proyecto se realiza un análisis del territorio en el que se utiliza el algoritmo para detectar las entidades de población que son objeto de estudio. Como resultado de esta fase, se obtiene
    a. Un listado de entidades a encuestar, con una lista de topónimos candidatos para cada una de ellas.
    b. Mapas de actuación en los que se muestra los límites de las entidades de población detectadas, así como los límites de las zonas de influencia de cada entidad.
    c. Ortofotos de cada una de las entidades, generadas de forma automática a través de una extensión de la aplicación SIG gvSIG, los cuales facilitaron el proceso de filtrado de entidades no rurales.

2. De forma paralela a esta primera tarea se realiza la definición de la ficha que determina la información a recoger en cada una de las entidades encuestadas.

3. Partiendo de las especificaciones de la ficha a cubrir en las visitas a las entidades de población se elabora una aplicación de introducción de datos para dispositivos móviles que permite:
    a. Recoger de datos de las entidades a través de una PDA.
    b. Generar ficheros XML con los datos recogidos en la PDA durante las encuestas. Estos ficheros serán enviados a un servidor para su posterior volcado a una base de datos centralizada.
    c. Leer de un fichero la lista de entidades que debe de encuestar, de forma que permita gestionar a cada grupo de trabajo las entidades encuestadas y pendientes de encuestar.

4. Con el material generado en las fases 1, 2 y 3, los equipos de recogida de datos salen “a campo” a encuestar las entidades detectadas utilizando PDAs.

5. La información recogida por los grupos de trabajo, y volcada a la BD centralizada, es publicada en Internet a través de una aplicación Web. Existe un control de acceso a la aplicación, y diferentes roles de acceso:
    a. Consulta: Podrá consultar la información encuestada y almacenada en la base de datos.
    b. Tutor de grupo de trabajo: Podrá consultar la información y además tendrá acceso a una herramienta que les permitirá realizar un análisis DAFO de las entidades.

6. En una última fase, se generan mapas e informes con los resultados obtenidos por los equipos de trabajo en las encuestas y por los tutores de grupo en los análisis.

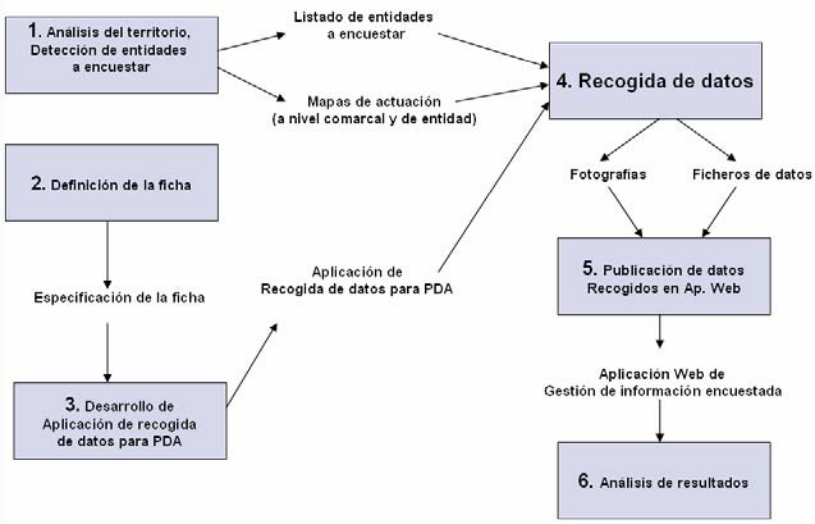

Figura 2: metodología del proyecto Diagnóstico del hábitat rural de Galicia

4 Conclusiones y trabajo futuro

En este artículo se ha descrito de forma detallada el algoritmo de detección de entidades de población del proyecto *Diagnosis del hábitat rural de Galicia*, se han evaluado diferentes alternativas para la detección de entidades, y se ha presentado el proyecto en el que se enmarca este trabajo.

Hasta el momento, se han completado las 4 primeras fases del proyecto Diagnosis del hábitat rural de Galicia, que ha abarcado 12 comarcas, 7.286 km2 y un total de 4.352 entidades. En este momento nos encontramos en la quinta fase del trabajo ultimando la aplicación web de explotación y terminando de recoger la información encuestada por los grupos de trabajo.

Como trabajo futuro se pretende analizar la restante superficie de Galicia, alcanzando las 53 comarcas, 29.574 km^2 de terreno analizado, y aproximadamente las 32.000 entidades encuestadas.

Agradecimientos

El proyecto *Diagnóstico del hábitat rural de Galicia* se está desarrollando a partir de un convenio entre la *Consellería de Vivenda e Solo* de la *Xunta de Galicia* y la Universidad de A Coruña. En él participan investigadores de la Escuela Técnica Superior de Arquitectura dirigidos por Xosé Lois Martinez, Plácido Lizancos, y Juan M. Doce.

El algoritmo de detección de entidades, así como las restantes aplicaciones informáticas del proyecto, han sido desarrolladas en el Laboratorio de Bases de Datos de la Facultad de Informática dirigido por Nieves R. Brisaboa.

Referencias

[1] Tammilehto-Luode, M. , Backer, L., Rogstad , L., Statistical Journal of the United Nations ECE 17 (2000) 109–117: Grid data and area delimitation by definition towards a better European territorial statistical system.

[2] Gallego, F.J.: Mapping rural/urban areas from population density grids. Institute for Environment and Sustainability, JRC, Ispra (Italy)

[3] Open Geospatial Consortium. Web Map Service Specification. Version 1.3. Retrieved August 2008 from: http://www.opengeospatial.org/standards/wms