Los SIG y el control de las operaciones de recogida de información estadística

Eduard Suñé Luis
Institut d'Estadística de Catalunya (IDESCAT) , esl@idescat.net

Resumen: *En el contexto de una futura operación exhaustiva tipo estadística de población y utilizando herramientas Open Source (JUMP y PostGIS) se ha desarrollado un prototipo para la monitorización de la cobertura y calidad de la información que se va recogiendo. Para ello se utilizan , entre otra , información del Institut Cartogràfic de Catalunya (ICC), catastro urbano y el registro de población como directorios base sobre los que realizar una simulación. En relación al control de calidad de la información recogida, se han implementado utilidades de análisis como cálculos de índices de autocorrelación espacial, localización de outliers espaciales, así como clustering y análisis de componentes principales. Finalmente se proponen evoluciones del prototipo en cuanto a arquitectura y funcionalidades.*

INTRODUCCIÓN.

La obtención de información en las operaciones exhaustivas tipo estadística de población implican, generalmente, la entrevista del ciudadano por parte del personal de la oficina estadística con el fin de cumplimentar un cuestionario previamente diseñado.

La captura y posterior tratamiento de los datos reseñados en el cuestionario serán el resultado final de esta costosa operación que sólo puede abordarse desde la táctica de descomposición en problemas más pequeños, es decir, compartimentando el territorio de tal manera que a un encuestador se le asigna una parte sobre el que tendrá que obtener los datos de forma exhaustiva (sección censal), apoyándose para ello, en unos directorios iniciales (registro de población o padrón).

Este encuestador está supervisado por un responsable de grupo que tiene asignada una parte mayor del territorio (un conjunto de secciones censales y por ende un conjunto de encuestadores). A su vez el responsable de grupo está supervisado, formando, todo esto, una estructura jerárquica asociada al territorio.

Es obvio que un sistema de información geográfico pueda desempeñar un papel fundamental en todo el proceso de control ya que

- Identifica las secciones censales (delimitación del espacio y asignación de responsabilidades)
- Puede asignar un cuestionario a zonas pequeñas (parcela catastral) mediante las direcciones postales, si los datos cartográficos y alfanuméricos son disponibles y de calidad
- Permiten la localización de anomalías de forma temprana a través de los datos de los cuestionarios, una vez capturados y agregados ya sea a nivel de sección censal, superior o incluso inferior

El punto más importante, el que determinaría la potencialidad real de estas herramientas en este contexto, es el de poder situar un cuestionario en el espacio, al nivel más detallado posible desde el punto de vista cartográfico.

Los directorios de personas (padrón o registro de población), de viviendas (padrón catastral), el resto de datos cartográficos y la información sobre los estados sucesivos de los cuestionarios constituyen las estructuras de datos necesarias para poder realizar un control minucioso de la operación de campo, respetando escrupulosamente, en todo momento, la salvaguarda del secreto estadístico.

En IDESCAT hemos desarrollado un prototipo, basándonos en conocidos proyectos Open Source (JUMP[1] y PostGIS[2]), con la finalidad de evaluar las posibilidades reales de un sistema de esta naturaleza, dotándolo de utilidades orientadas al análisis descriptivo de datos espaciales así como de clásicos métodos de análisis descriptivo multivariante.

CONTROL DE COBERTURA.

Las estructuras de datos del sistema pueden subdividirse en:

- Directorio de población inicial: hogares y personas. Cuestionarios iniciales.
- Límites administrativos del territorio: comarca, municipio, distrito y sección censal.
- Agentes de la oficina estadística: encuestadores, encargados de grupo, etc.
- Cartografía correspondiente al catastro urbano[3]: masas, parcelas, elementos lineales, etc.
- Información alfanumérica del catastro urbano (padrón catastral)

que en conjunto presentan el siguiente esquema entidad-relación simplificado:

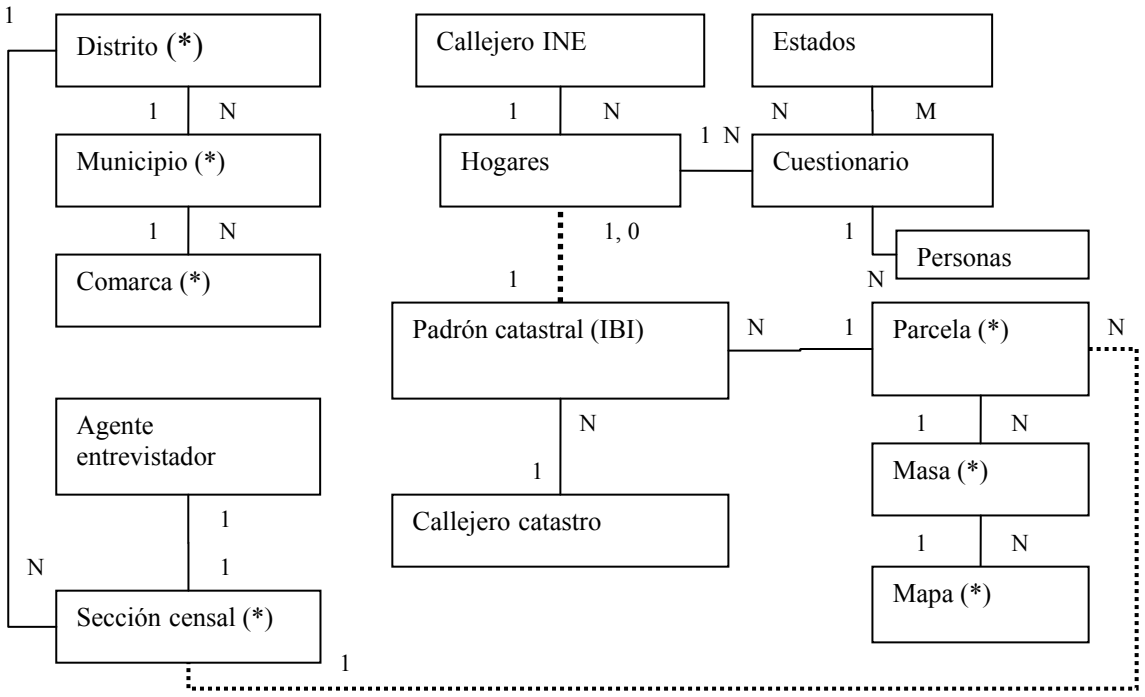


Figura 1: Esquema simplificado entidad-relación. Las entidades marcadas con un asterisco tinenen atributos geométricos y conforman las diferentes capas en el SIG.

En esta estructura de datos cabe destacar las relaciones entre las entidades Hogares - Padrón catastral y Sección censal – Parcela ya que provienen del tratamiento de las direcciones postales. La entidad cuestionario, mediante su relación con las entidades personas y estados representa la evolución de la operación de campo.

Los estados por los que puede pasar un cuestionario dependen del nivel de detalle que queramos tener de la operación, por ejemplo la sucesión de estados:

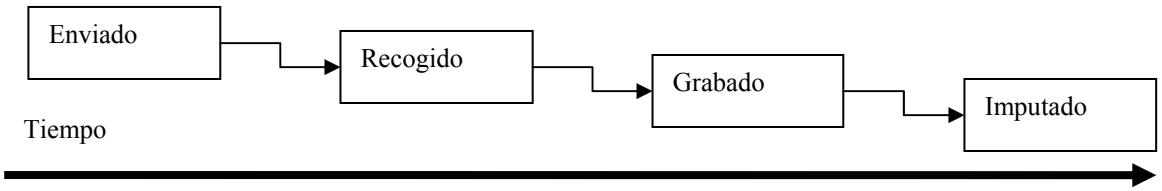


Figura 2: Una posible sucesión de estados. Cada cambio implica la modificación de las entidades correspondientes en la B.D.

obligaría a realizar las operaciones de modificación en la base de datos en los diferentes estadios lo cual puede representar un trabajo extra muy costoso (en la anterior sucesión podríamos eliminar el estado recogido ya que el estado

grabado lo implica necesariamente; no obstante ante la pérdida de cuestionarios no sabríamos si han sido recogidos o no o si por el contrario no han sido grabados accidentalmente).
El control del desarrollo de la operación de campo puede realizarse con la ayuda del SIG mediante la observación de mapas temáticos de los porcentajes de cuestionarios que están en cada uno de los estados anteriormente referidos, asociándolos al nivel territorial de interés de un usuario: así un encuestador estará interesado en la observación de los porcentajes de cuestionarios recogidos, grabados, etc., a nivel parcela dentro de una sección censal. A un encargado de grupo posiblemente le interesará observar los mapas a niveles superiores para poder realizar la supervisión de la que es responsable.

La aplicación que hemos desarrollado tiene actualmente una arquitectura en dos capas. El back-end corresponde a una base de datos postGIS sobre la que se han implementado las estructuras descritas en el apartado anterior. Las tablas que disponen de columnas de tipo "geometría" han sido indexadas via R-tree [4], esquema de indexación espacial soportado por postGIS. De esta manera la obtención de las capas catastrales para una sección censal dada puede realizarse a través de un query geo-espacial del estilo:

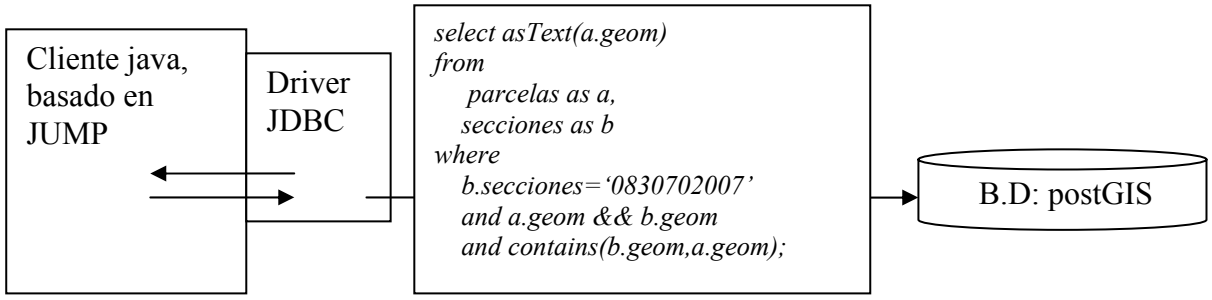


Figura 3: Arquitectura actual de la aplicación y un query geo-espacial ejemplo en postGIS. Nótese que no es necesario que la capa parcela tenga asignados códigos de sección

Por otro lado, el cliente java está basado en el conocido proyecto JUMP utilizando el mecanismo que proporciona su arquitectura para realizar las extensiones pertinentes. Este mecanismo se basa en el desarrollo de clases que implementan unas interfícies conocidas:

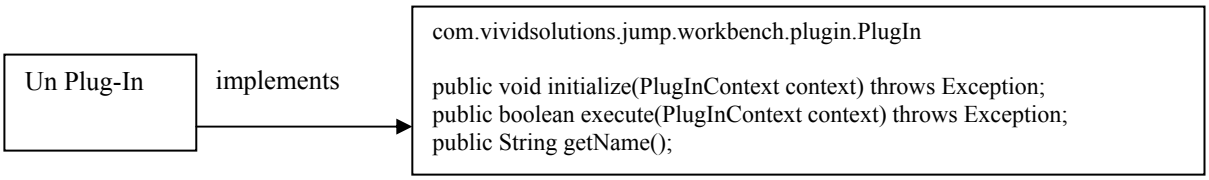


Figura 4: La especificacion JUMP relativa a plug-ins

Tanto en los miembros *initialize* como *execute* se pasan referencias a instancias de la clase *PlugInContext* que a su vez contiene referencias a los objetos que la aplicación está manipulando como ventanas, capas etc. El miembro *execute* se dispara como acción de un elemento del menú que es adaptable a las necesidades de una aplicación concreta.

Para la realización del control de la operación de campo, se han desarrollado los plug-Ins para la obtención de los mapas correspondientes a las secciones censales, parcelas, masas etc, como resultado de la entrada de un código de sección censal. Además, gracias a que JUMP dispone de un cliente WMS, es posible obtener ortofotos y otras capas de los servidores públicos del Institut Cartogràfic de Catalunya (ICC) [5] permitiendo así un mejor conocimiento del área asignada a un encuestador

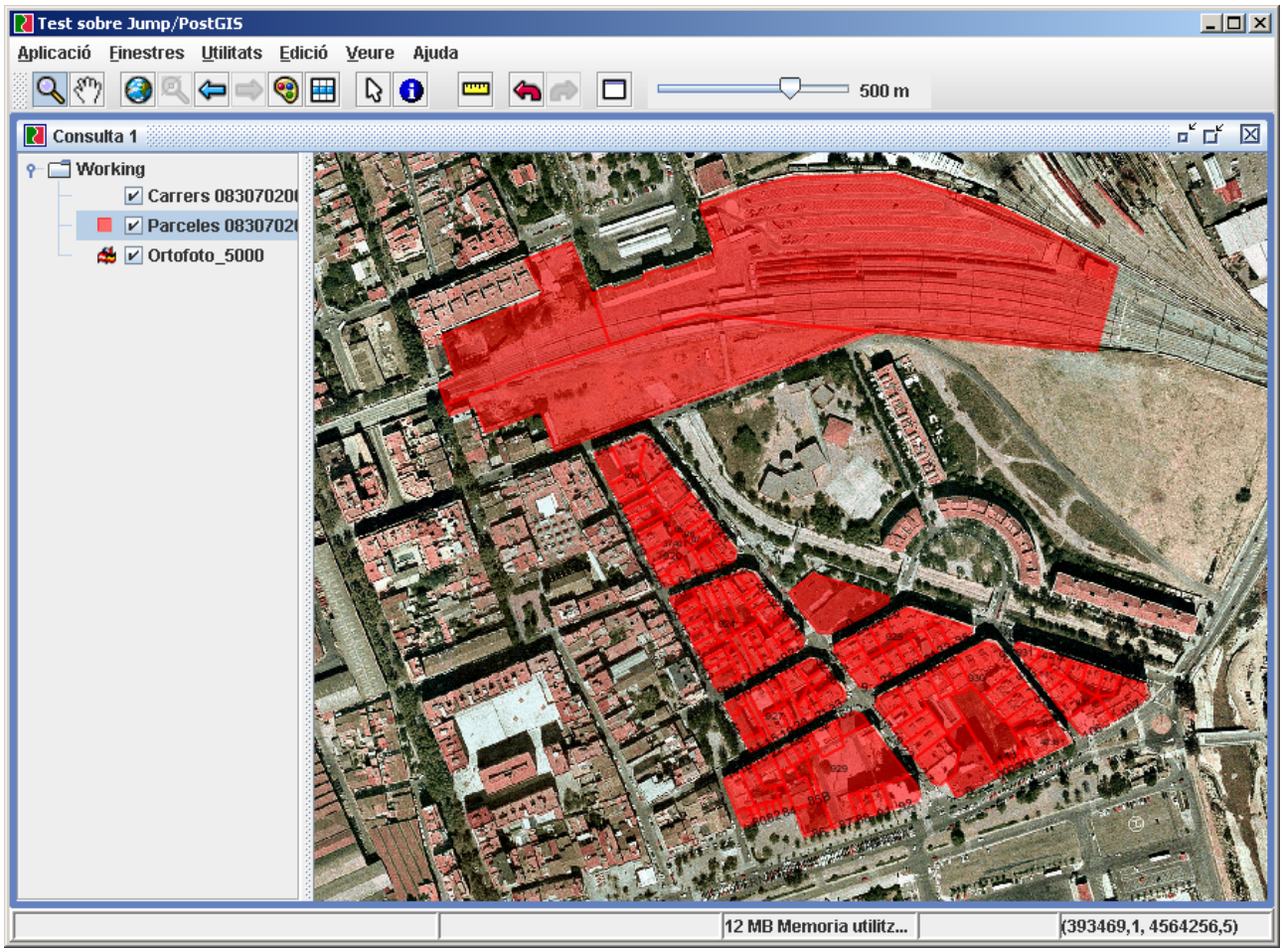


Figura 5: Mapa correspondiente a las parcelas de una sección censal de Vilanova i la Geltrú. De fondo la capa generada por el servidor WMS del ICC.

Al ejecutar este plugin no solamente obtenemos de la base de datos la información asociada a las capas sinó que se evaluan los siguientes datos:

- Número de hogares
- Número de personas
- % de cuestionarios enviados
- % de cuestionarios recogidos
- % de cuestionarios grabados
- % de cuestionarios imputados

agregados al nivel más bajo que corresponda (en este caso a nivel de parcela catastral), y que representa el estado de la operación de campo en ese momento.

La representación de mapas temáticos de esas variables proporcionan una fotografía del estado de la operación de campo en esa área de interés:

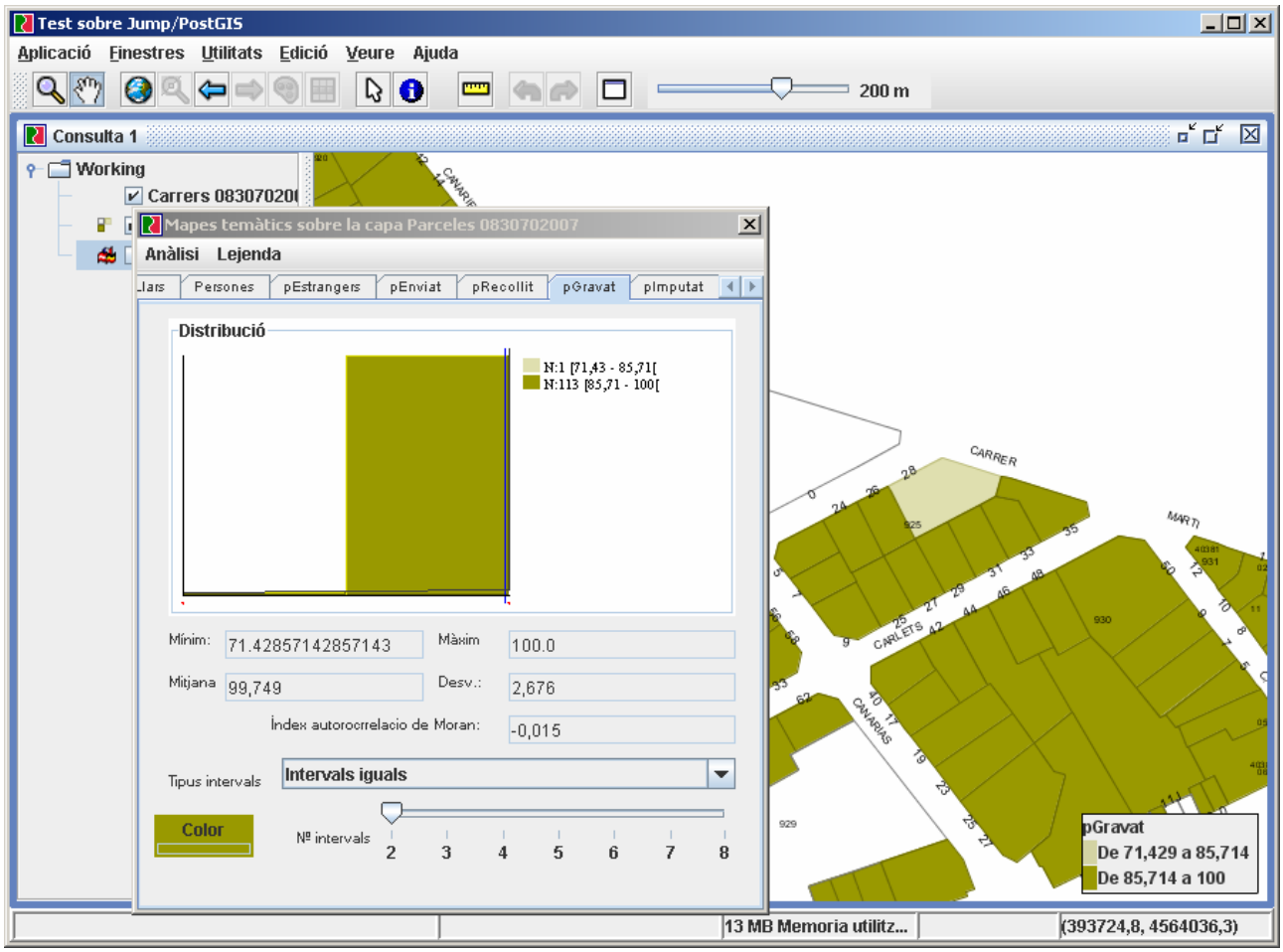
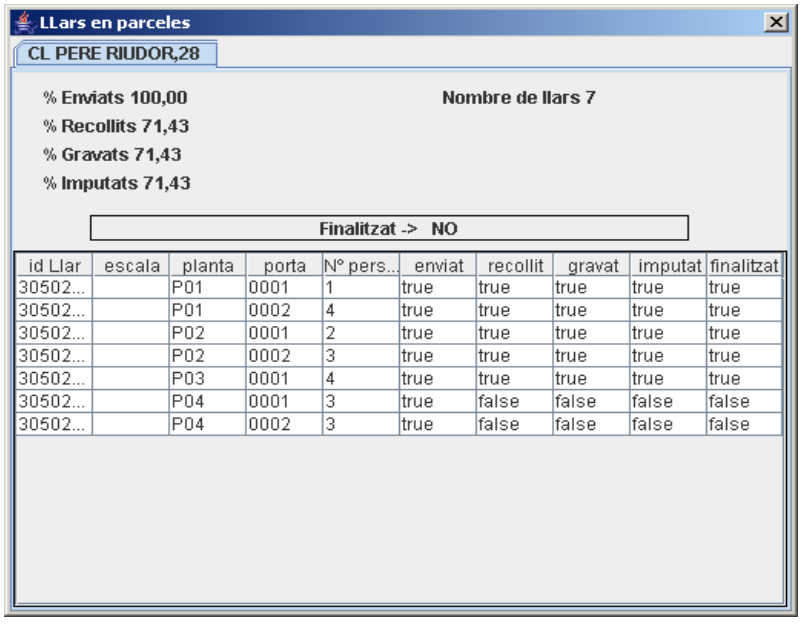


Figura 6: Distribución (simulada) del porcentaje de cuestionarios grabados en una sección censal de Vilanova i la Geltrú. Nótese que el agente encuestador puede conocer en que lugar se descubre la falta de cuestionarios grabados. Se ha omitido la capa ortofoto proporcionada por el WMS del ICC.

Al seleccionar una parcela se puede obtener información detallada de los hogares y estado de los cuestionarios:



id Llar	escala	planta	porta	Nº pers...	enviat	recollit	gravat	imputat	finalitzat
30502...		P01	0001	1	true	true	true	true	true
30502...		P01	0002	4	true	true	true	true	true
30502...		P02	0001	2	true	true	true	true	true
30502...		P02	0002	3	true	true	true	true	true
30502...		P03	0001	4	true	true	true	true	true
30502...		P04	0001	3	true	false	false	false	false
30502...		P04	0002	3	true	false	false	false	false

Figura 7: Hogares en la parcela, número de personas y estado de los cuestionarios

Otros casos de usos implementados son el acceso a los datos del cuestionario (simulado) para un hogar determinado. Esta opción abre la puerta a poder editar los datos del cuestionario directamente desde la aplicación, caso especialmente útil para encuestas no exhaustivas en las que el agente disponga de un portátil y para el que se le prepararía una aplicación local de este estilo.

Esta aplicación ha sido desarrollada con una arquitectura en dos capas, acceso a postGIS via drivers java de tipo IV (thin drivers) y su despliegue se realiza con Java WebStart. La migración a arquitecturas más escalables de tres capas son viables gracias a las implementaciones de referencia de CachedRowSet [6] con el concurso de un servidor http i un motor de servlets con una máquina virtual nivel 1.5

CONTROL DE CALIDAD DE LA INFORMACIÓN RECOGIDA.

Una vez grabada la información se procede a su análisis para verificar su calidad, nivel de respuesta etc. Normalmente este análisis se realiza observando la distribución de valores así como los resultados obtenidos de un conjunto de tabulaciones cruzadas. Otra forma de verificar la bondad de los resultados es observar como son las distribuciones en relación al espacio. Para ello, la aplicación de control implementa la generación de los típicos mapas temáticos, con las siguientes opciones:

- Intérvalos iguales
- Mapas asociados a percentiles
- Dos intérvalos por encima /debajo de la media
- Tres intérvalos. Central una desviación o central dos desviaciones
- Cluster mediante KMeans
- Por encima/debajo de cero, si hay cambio de signo

Para datos referentes al Censo de población 2001 para la comarca del Barcelonés y para el porcentaje de Licenciados y doctores a nivel de sección censal obtendríamos :

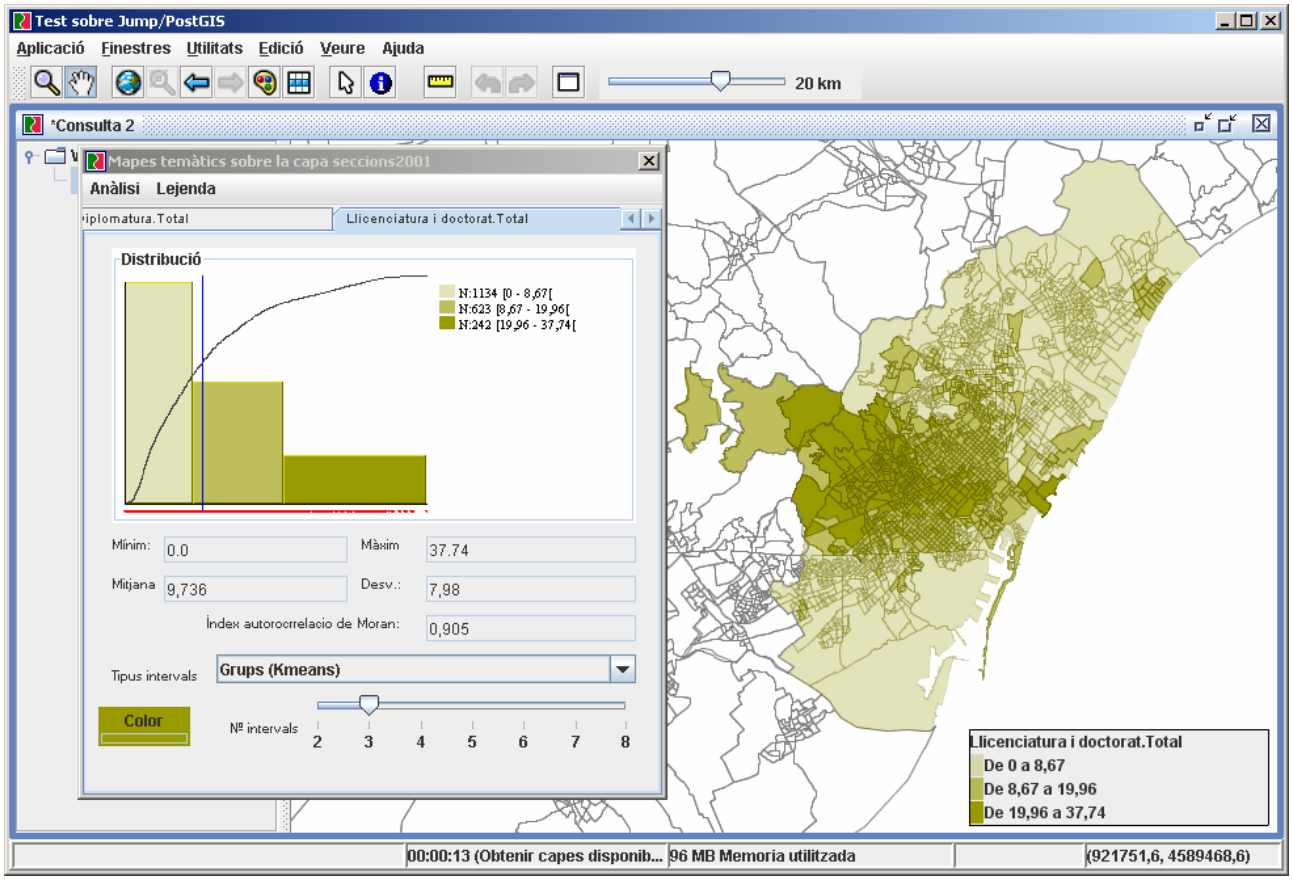


Figura 8: Mapa temático del % de Licenciatura y doctorado. Comarca del Barcelonés

La representación en la fig. 8 se corresponde con el método de generación de clusters KMeans utilizando la implementación realizada en el proyecto Weka [7].

Como puede observarse existe una fuerte autocorrelación espacial (las zonas de elevado % de licenciados y doctores son compactas) tal como indica el índice de autocorrelación de Moran [8] definido por:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

donde

n es el número de áreas
yi es el valor observado en el elemento i -ésimo
wij es la distancia entre los elementos i-esimo y j-ésimo
y es el valor medio de la variable

Este índice, cuyos valores están comprendidos entre -1 y +1, mide la autocorrelación espacial de una variable, siendo el valor 0 indicativo de ausencia de autocorrelación. El elemento wij representa la distancia entre i y j siendo su valor dependiente del esquema de medición de distancias empleado. En nuestro caso, por razones de eficiencia,

wij = 0 si los polígonos i y j no se tocan
wij = 1 si los polígonos i y j se tocan

y se evalúa mediante el api JTS [1], utilizado internamente por JUMP, gracias a que cumple con las especificaciones SFSQL de la OGC del lado cliente.

La simple observación de valores de la distribución, o la información del grado de autocorrelación espacial indicarían de forma global la bondad de los datos (por ejemplo, una elevada autocorrelación espacial para la edad sería muy sospechoso).

No obstante para nuestros objetivos es más importante la detección de outliers espaciales, es decir zonas con valores anómalos en relación a sus vecinos (que no serían fácilmente detectables con los mapas anteriores o con las tabulaciones pertinentes). La representación simultánea de los valores observados frente al promedio ponderado del de sus vecinos (gráfico de Moran) facilita la detección tanto de outliers como de outliers espaciales.

En la implementación realizada para este gráfico se presentan la nube de puntos, la recta cuya pendiente coincide con el índice de autocorrelación global de Moran, un área gris que corresponde a valores dentro del intervalo central de la distribución de distancias entre el valor observado y los promedios espaciales, así como dos líneas marcadas en magenta que señalan los valores centrales de la distribución.

Los puntos en el gráfico quedan agrupados en cuatro cuadrantes :

- L-H valores bajos y vecinos con valores altos
- H-H valores altos con vecinos con valores altos
- H-L valores altos con vecinos de valores bajos
- L-L valores bajos con vecinos de valores bajos

Los valores que quedan fuera de las líneas de color magenta pueden considerarse outliers en el sentido clásico de término. Los valores que quedan fuera del área gris pueden considerarse outliers espaciales, siendo los situados en la zona superior los valores que son más bajos que los de su entorno y los de la zona inferior valores superiores a su entorno.

El usuario, gracias a que esta ventana se ha registrado como un listener de la ventana de representación de los mapas, puede seleccionar una geometría (en este caso sección censal) y observar ese punto en el diagrama de Moran y al contrario, puede seleccionar una área en el diagrama y observar a que partes del territorio les corresponde :

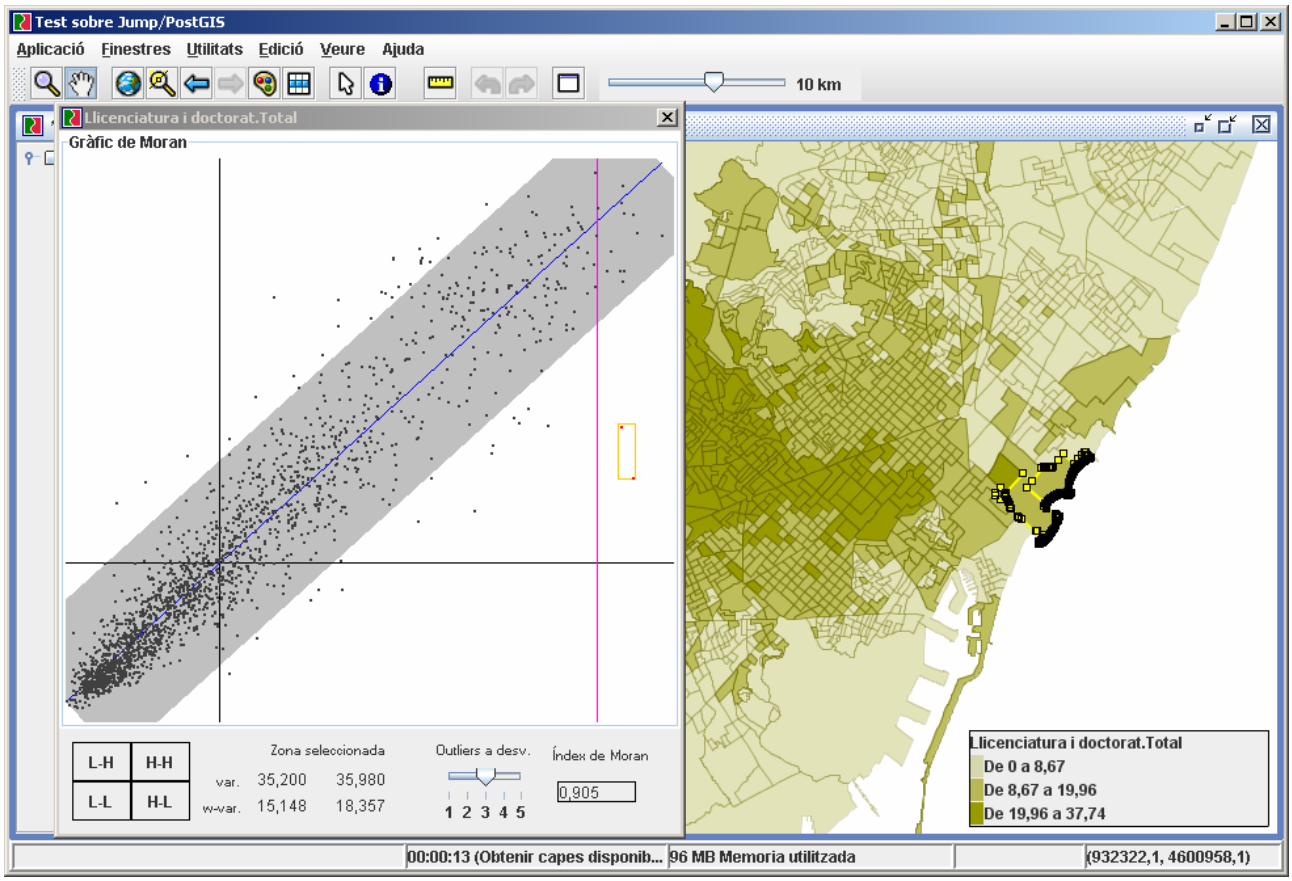


Figura 9: La selección desde el gráfico de Moran produce la selección en el mapa y al contrario. Obsérvese que los puntos seleccionados pueden considerarse tanto outliers como outliers espaciales.

Finalmente, para poder analizar mejor los datos desde una perspectiva multivariante, se ha implementado el análisis de componentes principales (ACP) utilizando para ello el api de tratamiento de matrices Jama [9]. El ACP es una técnica multivariante que trata de obtener un conjunto reducido de factores ortogonales, combinación lineal de las variables originales, de tal manera que no exista una pérdida elevada de información. Puede ser útil en el contexto del control de calidad ya que ilustra con claridad las interrelaciones entre variables y permite localizar outliers (espaciales y no espaciales) desde una perspectiva multivariante. Para ilustrar el análisis se presentan los resultados utilizando los valores correspondientes al censo de población 2001 de las variables:

% Licenciados y doctores
Tasa de paro
% Técnicos, profesionales y científicos
% Trabajadores cualificados de la industria y construcción
% Trabajadores no cualificados

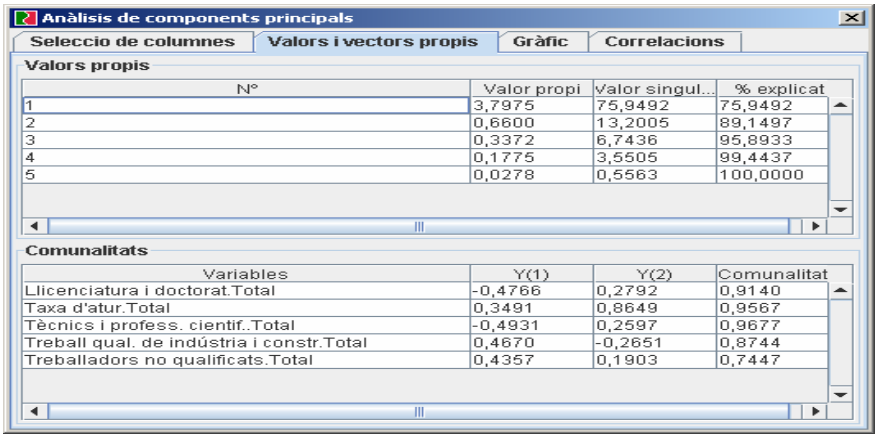


Figura 10: Resumen del ACP. Vectores y valores propios

Vemos que con solo dos factores explicamos el 89% de la varianza total. La interpretación de los factores pasa por observar las contribuciones de las variables originales: el segundo factor va en la dirección de la tasa de paro, mientras que el primero es una combinación lineal de todas las variables pero que sitúa en direcciones contrarias los trabajos más técnicos de los manuales siendo los valores bajos en el factor, altos en el % de técnicos, profesionales y científicos. Los puntos situados en valores altos del primer y segundo factor se corresponderán a valores altos de tasa de paro.

En el gráfico de los factores del ACP pueden representarse tanto los individuos (en este caso secciones censales) como las direcciones de las variables originales en relación a los factores. De nuevo en la ventana del gráfico es posible seleccionar puntos y localizar en el mapa a qué elementos corresponden y al contrario :

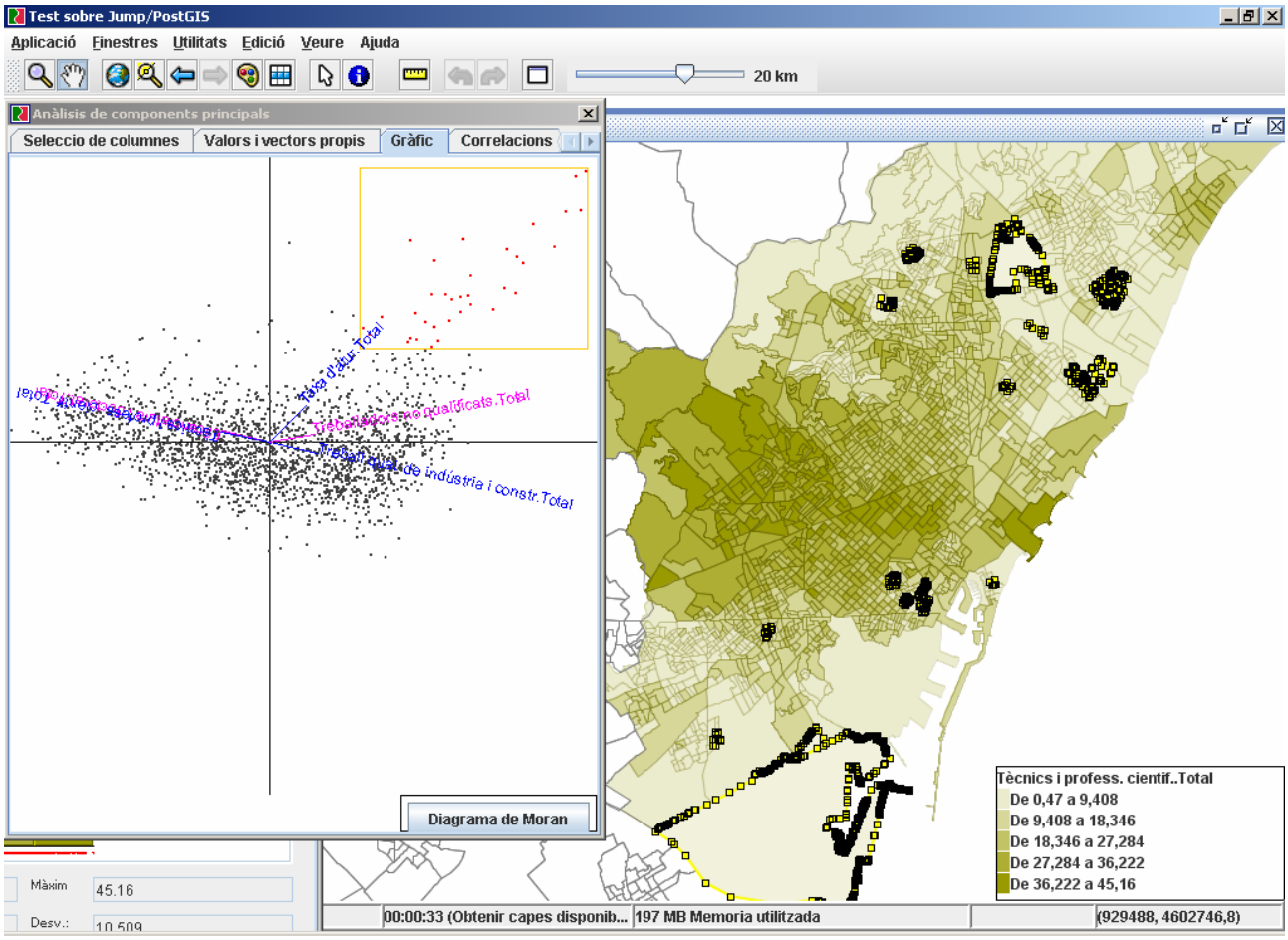


Figura 11: Gráfico del ACP. Los elementos seleccionados están en la dirección de Tasa de paro y % de Trabajadores no cualificados. Se corresponden con valores bajos de % de Licenciados y doctores.

Véase además que la tasa de paro, contribución importante al segundo factor, es casi ortogonal con el primero, aunque está más en la dirección del % de trabajadores no cualificados. Estos resultados, que podríamos considerar obvios a nivel global, indican que nuestros datos tienen (globalmente) una calidad aceptable.

Sobre las coordenadas en esos factores de los individuos es posible calcular el índice de autocorrelación espacial de Moran y construir el gráfico para poder situar outliers espaciales y no espaciales en el espacio de factores del ACP, análogamente a los cálculos realizados sobre una variable.

Un índice de autocorrelación de Moran próximo a la unidad indicara que hay zonas compactas relativas a los valores del factor considerado. La localización de outliers espaciales en relación a los factores del ACP puede revelar anomalías en relación al conjunto de variables que determinan ese factor. En la siguiente figura se presenta el gráfico de Moran relativo al primer factor del ACP :

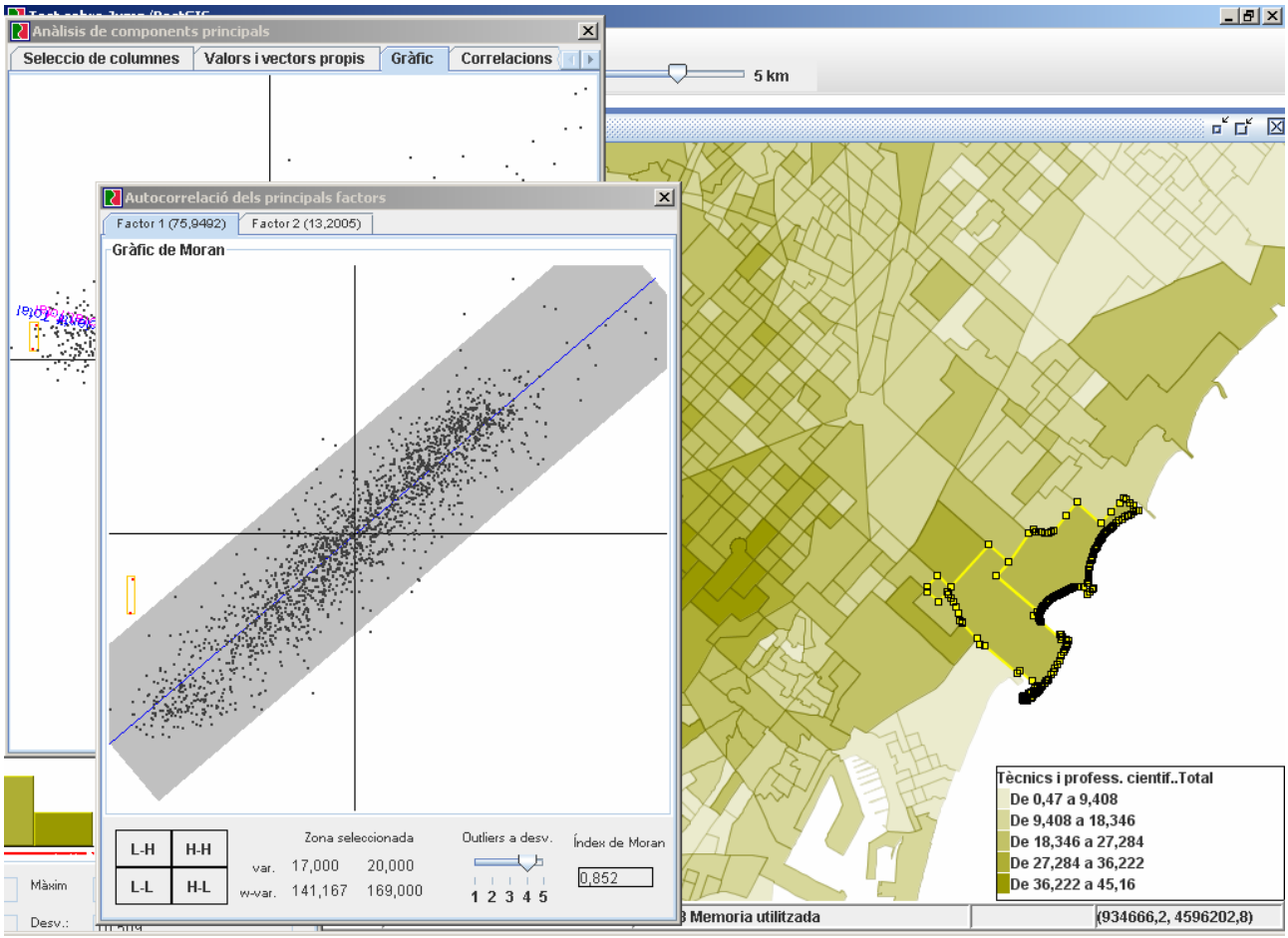


Figura 12: Diagrama de Moran sobre el primer factor del ACP.

Obsérvese, en la fig. 12, que los elementos seleccionados son outliers espaciales situados en valores bajos del primer factor que pueden interpretarse como elevado % de Licenciados, elevado % de técnicos, profesionales y científicos y bajo % de trabajadores no cualificados. Nuestro conocimiento de los datos y el territorio nos indica que estos outliers espaciales no lo son por una mala calidad de los datos, no obstante los métodos implementados permiten localizarlos fácilmente.

FUTUROS DESARROLLOS

En cuanto a la arquitectura de la aplicación tenemos previsto migrar de una arquitectura en dos capas a tres, utilizando un servidor http intermedio y las implementaciones de referencia de CachedRowSet del Api java 1.5. Se implementarán utilidades como la generación de informes del estado de la operación y se derivará una versión específica para el análisis de datos espaciales. Se implementará el análisis de correspondencias simples para poder analizar tablas de contingencia de forma equivalente al ACP, con la finalidad observar la relación entre las categorías columnas (variables) y filas (espacio) y realizar los cálculos de coeficientes de autocorrelación en relación a los ejes factoriales obtenidos.

REFERENCIAS

1. Vivids solutions website. http://www.vividsolutions.com/
2. PostGIS website. http://postgis.refractions.net/
3. Dirección General del Catastro website. http://www.catastro.minhac.es/
4. Guttman, A., 1984, R-trees: A dynamic index structure for spatial searching. Procedings of the ACM SIGMOD International Conference on Management of Data.
5. Institut Catogràfic de Catalunya ICC website. http://www.icc.es/
6. Java Technology website. http://java.sun.com/
7. Weka : Data Mining Software in Java website. http://www.cs.waikato.ac.nz/ml/weka/
8. Anselin, L. (1994). "Local indicators of spatial association - LISA", Techn. Rep. 9331. Regional Research Institute, West Virginia University, Morgantown WV 26506-6825 USA.
9. JAMA : A Java Matrix Package website. http://math.nist.gov/javanumerics/jama/