Algunos desarrollos basados en JUMP para el apoyo logístico y control de calidad en las operaciones de recogida de información estadística.

E. Suñé Luis [(1)]

[(1)] Subdirección de Producción Estadística. Institut d'Estadística de Catalunya. Via Laietana 58, 08003, Barcelona, esl@idescat.net.

RESUMEN

Las operaciones de recogida de información estadística de tipo exhaustivo presentan grandes desafíos logísticos. Con la utilización de los SIG la gestión de la información asociada se simplifica y el control de calidad de la información recogida puede ser más estricto.

Con esta doble finalidad, el control de la operación de campo y de calidad de resultados, se han desarrollado un conjunto de plugins JUMP, que van desde la elaboración de mapas de estados a la utilización de técnicas multivariantes como el análisis de correspondencias, utilizando en su desarrollo conocidos proyectos open source como WEKA, JAMA, así como gráficos 3D. Estos desarrollos conforman, junto con una base de datos implementada en postGIS, un sistema de información para el control logístico y de calidad de una futura operación de campo exhaustiva.

Palabras clave: *SIG, JUMP, censos, análisis de correspondencias, clusters*

ABSTRACT

Exhaustive statistical information-gathering operations pose major logistical challenges. By using GISs, managing the associated information becomes simpler, and monitoring the quality control of the information gathered can be stricter.

With this two-fold purpose in mind, namely monitoring the field operation and quality of the results, a series of JUMP plug-ins have been developed, ranging from the generation of maps of states to the use of multivariant techniques such as correspondence analyses. To develop them, well-known open source projects such as WEKA and JAMA have been used, along with 3D graphics. These developments, along with a data base implemented postGIS, make up an information system for logistical and quality control for a future exhaustive field operation.

Key words: *GIS, JUMP, censuses, correspondence analysis, clusters.*

1. INTRODUCCIÓN

Los sistemas de información geográficos pueden desempeñar un papel importante en las operaciones presenciales de recogida de información estadística. En estas operaciones se debe asegurar el máximo de calidad y exhaustividad, tanto en unos censos, cuyo objetivo es toda la población, como en encuestas, cuyo objetivo es una parte, convenientemente seleccionada mediante métodos de muestreo.

Estas operaciones presentan grandes desafíos de tipo logístico sobre todo en encuestas con grandes tamaños muestrales o, obviamente, en censos. En estos casos se procede a la segmentación de los directorios iniciales mediante una cierta partición del espacio, asignándose así responsabilidades a los grupos de personas que van a realizar el trabajo de campo. Que el grupo de personas que realiza la operación disponga de una aplicación SIG que le informe del estado de recogida de los cuestionarios asociados a su área puede tener una gran trascendencia en cuanto a asegurar los máximos niveles de exhaustividad.

Los cuestionarios se irán recogiendo y la información resultante se someterá a procesos de codificación automática, validación e imputación. Se obtendrán unos resultados que deberían ser válidos en los diferentes niveles geográficos para los que la operación sea significativa, que para un censo va desde todo el territorio hasta el nivel de sección censal. Aquí los SIG también pueden desempeñar un gran papel al poder revelar posibles anomalías de las distribuciones obtenidas para los diferentes niveles de agregación territorial.

En Idescat hemos desarrollado una serie de prototipos, basados en el conocido proyecto open source JUMP[1], con el fin de realizar el control logístico y de calidad al que nos referimos.

2. JUMP. MECANISMO DE EXTENSIÓN

JUMP es un SIG de escritorio, tipo MDI basado en Swing y java puro, desarrollado por Vivid Solutions, que permite la visualización y manipulación de datos espaciales. Su mecanismo más sencillo y directo de extensión se basa en el desarrollo de plugins, clases que deben implementar la interficie *com.vividsolutions.jump.workbench.plugin.PlugIn* definida por las siguientes funciones miembro:

public void initialize(PlugInContext context) throws Exception;
public boolean execute(PlugInContext context) throws Exception;
public String getName();

Tanto en *initialize* como en *execute* el framework proporciona una referencia a una instancia de la clase *vividsolutions.jump.workbench.plugin.PlugInContext* que permite el acceso a toda la información que la aplicación esta manipulando, desde referencias a ventanas, datos espaciales, ítems seleccionados etc. El miembro *getName* devuelve un string que será utilizado por el elemento de menú que el framework creará para nosotros y cuya selección disparará *execute*.

En principio todo lo que debemos hacer como usuarios del API es desarrollar el plugin conforme a las especificaciones y desplegarlo en el directorio adecuado. Existe una muy buena documentación al respecto en el web site de JUMP.

La clase *PluginContext* tiene, entre otros, los siguientes miembros

public WorkbenchFrame getWorkbenchFrame()
public JInternalFrame getActiveInternalFrame()
public LayerViewPanel getLayerViewPanel()
public Layer getSelectedLayer(int i)

con los que podemos obtener las referencias a la ventana principal de la aplicación, la ventana interna activa, el panel donde se dibujan las diferentas capas de la ventana interna activa y la capa seleccionada (ver figura 1).

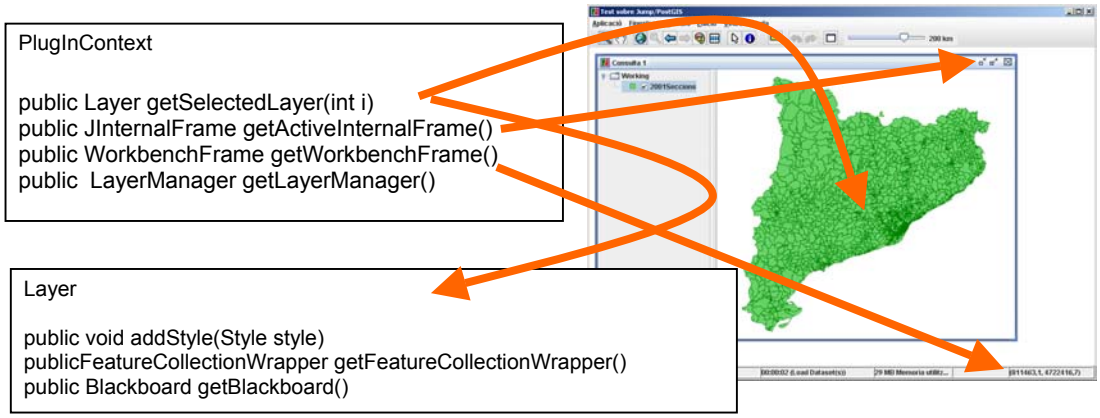

Figura 1. Las clases PluginContext y Layer

La clase *Layer* definida en el package *com.vividsolutions.jump.workbench.model* tiene, entre otros, el miembro público:

public FeatureCollectionWrapper getFeatureCollectionWrapper()

mediante los miembros públicos de *FeatureCollectionWraper* podemos obtener una colección de *Features* y los datos asociados:

public List getFeatures()
public FeatureSchema getFeatureSchema()

Con esto estamos en condiciones de poder modificar las capas que la aplicación manejará, tanto a nivel de sus geometrías como de los atributos definidos en el *FeatureSchema* asociado.

Fundamentalmente, los plugins que hemos desarrollado modifican el conjunto de atributos alfanuméricos de la lista de geometrías para realizar posteriormente cálculos muy específicos no implementados en el proyecto inicial. También tratan los eventos que se producen al realizar selecciones en el mapa tal como se expone más adelante.

3. APOYO LOGÍSTICO EN UNA OPERACIÓN CENSAL

Como indicábamos en la introducción el seguimiento de una operación de campo mediante SIG, sobre todo en el caso de una operación censal, puede ayudar a asegurar niveles altos de exhaustividad. Pongamos por caso el último censo de población y vivienda de 2001 realizado por el INE: la operación se llevo a cabo con el concurso de más de 40.000 agentes censales que recorrieron el territorio y recogieron los cuestionarios de 13 millones de hogares y de unos 40 millones de personas.

Es evidente que una operación así sólo puede realizarse en base a la segmentación del territorio y la asignación de una pequeña parte (una sección censal) a cada agente. Esa segmentación lo es también de los directorios sobre los que se apoyaría la operación, que en el caso de un censo podrían ser en principio el propio padrón de habitantes e información relativa al conjunto de inmuebles donde la población podría residir. El vehículo físico de la operación es un conjunto de cuestionarios (de hogar, de personas) que deberían estar correctamente georreferenciados para poder hacerlo con la información resultante. Para ello disponemos de las direcciones postales del directorio de partida: el padrón de habitantes.

Para poder utilizar un SIG en la operación necesitaríamos información espacial que esté relacionada con esos cuestionarios. La información espacial y alfanumérica del catastro urbano es el candidato más obvio para ese soporte; para ello deberíamos crear una base de datos como la de figura 2, de tal manera que cada cuestionario estuviese relacionado con la entidad espacial *parcela catastral*

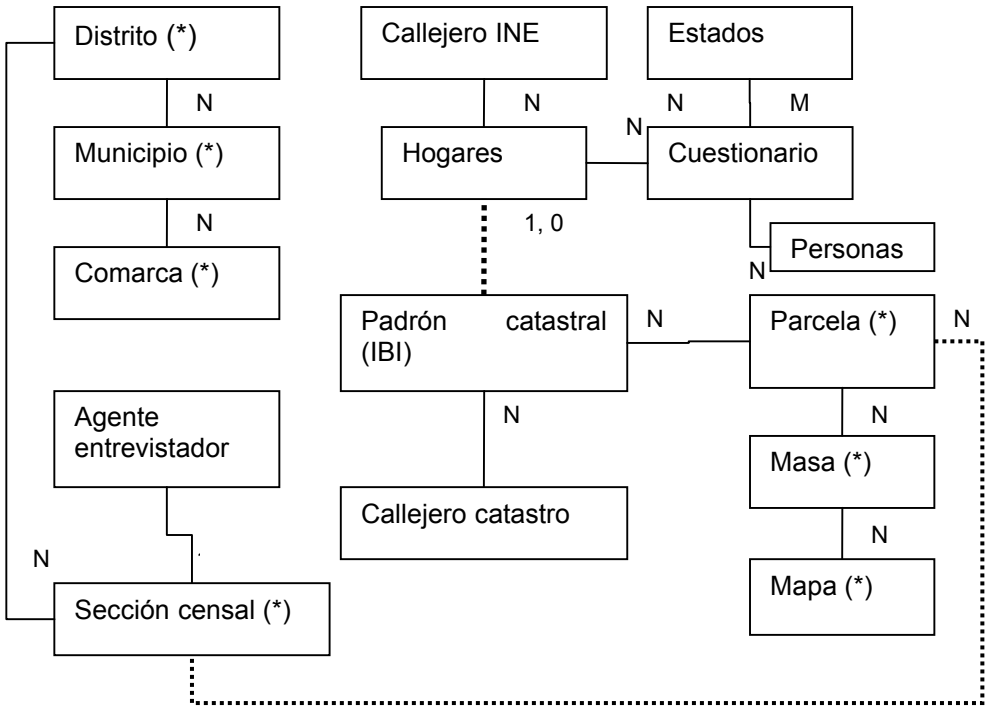

Figura 2.: Esquema simplificado entidad-relación. Las entidades marcadas con un asterisco tienen atributos geométricos y conforman las diferentes capas.

No obstante, el INE y la Dirección general del Catastro utilizan en sus respectivas bases de datos un conjunto de códigos diferentes. Sería, pues, necesaria la creación de tablas de correspondencias como ya se ha planteado en otros proyectos como Geopista o CartoCiudad.

3.1. Seguimiento de estados

Suponiendo que pueda establecerse una aplicación inyectiva entre el conjunto de hogares (y por tanto de cuestionarios) y el de parcelas, el seguimiento de la operación via SIG puede realizarse. Para ello es necesario mantener, a nivel de cuestionario, los estados por los que puede pasar, como los indicados en la figura 3.

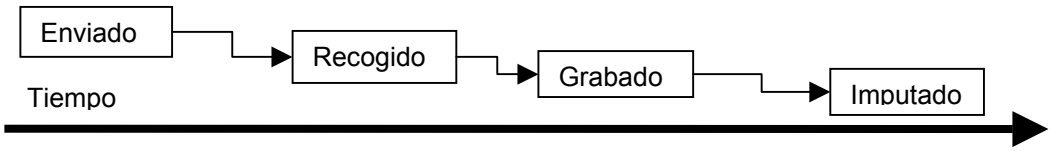

Figura 3. Una posible sucesión de estados

A nivel de parcela catastral se podrían obtener indicadores sobre el conjunto de cuestionarios y estados y obtener así los típicos mapas temáticos, en este caso a nivel de parcela y en relación al porcentaje de cuestionarios en cada estado. Los plugins JUMP necesarios para realizar este trabajo han sido desarrollados y la información de base se ha almacenado en la base de datos open source postGIS[2]. Con esto un agente censal tendría a la vista el resultado de su trabajo, tal como muestra la figura 4

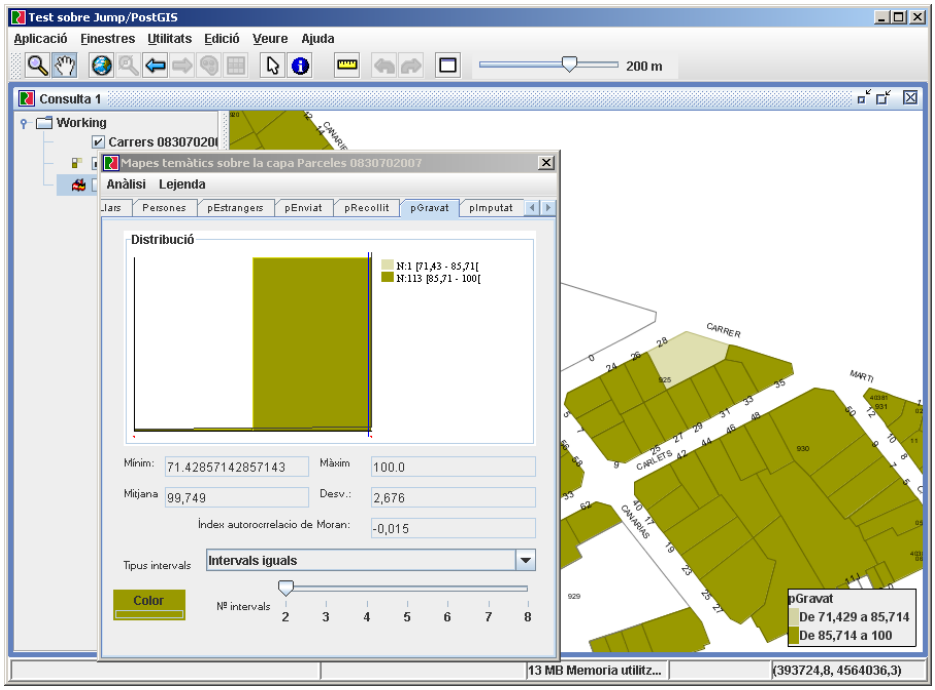

Figura 4. Distribución simulada del porcentaje de cuestionarios grabados en una sección de Vilanova i la Geltrú. Nótese que el agente puede conocer dónde se producen anomalías en la operación.

4. CONTROL DE CALIDAD

La información resultante de la grabación de los cuestionarios será tratada convenientemente. Se procederá a la codificación automática de literales, se la someterá al conjunto más exhaustivo posible de normas de validación y finalmente se procederá a la imputación de los datos faltantes o no correctos. Todo el proceso deberá someterse a estrictos controles de calidad.

En nuestro caso, estamos más interesados en observar como son las distribuciones espaciales de los datos recogidos, o en alguna de las fases de su tratamiento. Este control deberá realizarse en cualquiera de los niveles territoriales sobre los que la operación tenga sentido, es decir, para un censo, las distribuciones espaciales de las cuestiones reseñadas deberían ser razonables a niveles que van desde la comarca, hasta la sección censal pasando por municipio y distrito.

4.1. Detección de outliers en la distribución y en el espacio

En la figura 5 puede observarse la pregunta relativa a la actividad económica de los ocupados tal como se hizo en el censo de población del 2001:

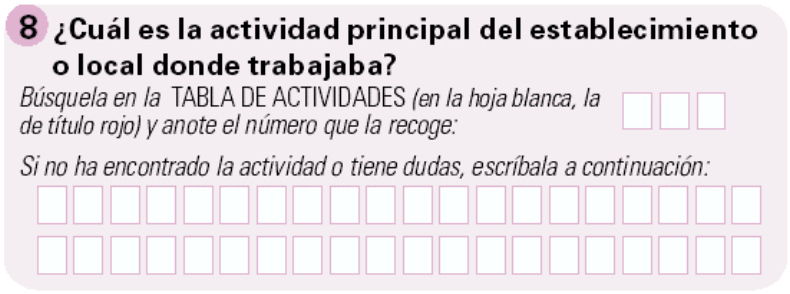

Figura 5. La pregunta de actividad del censo del 2001

Para esta pregunta, una vez realizados los procesos de codificación automática, validación e imputación obtendríamos por agregación en los diferentes niveles territoriales unas distribuciones que deberían ser aceptables. El SIG debería permitirnos detectar fácilmente anomalías en esas distribuciones, tanto en un sentido global (outliers en la distribución) como local (outliers espaciales).

Para esta pregunta existen más de 200 categorías diferentes y si estuviéramos analizando las distribuciones para el caso de las secciones censales de la Comarca del Barcelonés, unas 1996, la tarea sería ingente. Afortunadamente, existen técnicas de análisis de datos multivariantes que pueden ayudarnos a reducir la dimensionalidad del problema y con el concurso del SIG a caracterizar el comportamiento de nuestros elementos y localizar esas anomalías.

4.1.1. El análisis de correspondencias simples

Como quiera que muchas de nuestras variables son cualitativas, como la actividad, y una vez agregada la información en los diferentes niveles territoriales obtenemos tablas de contingencia, el análisis de correspondencias simples (AC) se perfila como una técnica muy adecuada.

El AC, ideado en los años 70 por el profesor Benzécri [3], considera las I filas (o columnas) elementos en un espacio de I-1 dimensiones con una distancia definida por

$$d^2(i,i') = \sum_{j=1}^{J} \left\{ \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \frac{f_{i'j}}{f_{i'.}\sqrt{f_{.j}}} \right\}^2 \quad (1)$$

donde

$f_{ij} = k_{ij} / K$ (2) frecuencia relativa del elemento (i,j)

$K = \sum_{i}^{I} \sum_{j}^{J} k_{ij}$ (3) número total de observaciones

$f_{i.} = \sum_{j}^{J} f_{ij}$ (4) frecuencia relativa marginal de la fila i-ésima

$f_{.j} = \sum_{i}^{I} f_{ij}$ (5) frecuencia relativa marginal de la columna j-ésima

para esta nube de puntos es posible encontrar un nuevo sistema de referencia que maximice la inercia explicada por cada uno de los factores y para ello es necesario diagonalizar la matriz

$$T = X^t X \quad (6)$$

donde

$$x(i,j) = \frac{\left(f_{ij} - f_{i.} f_{.j}\right)}{\sqrt{f_{i.}\ f_{.j}}} \quad (7)$$

Los valores propios asociados a la matriz T son proporcionales a la inercia explicada por cada factor. Usualmente se retiene un subespacio de dimensión tal que explique un cierto porcentaje de la inercia total. La utilización de esta técnica a problemas como los descritos anteriormente permite la reducción de dimensionalidad y la observación de posibles anomalías tal como veremos más adelante.

4.1.2. Autocorrelación espacial. Diagrama de Moran

En nuestro caso, aplicaremos esta técnica a las variables de un censo agregadas para niveles geográficos de tal manera que nuestros puntos columna sean las categorías de esa variable y nuestros puntos fila sean unidades territoriales. Una vez aplicado el AC, nuestros puntos fila podrán ser descritos en un subespacio de R^n con mayor o menor pérdida de información pero está claro que el concepto de distancia entre esos elementos fila es aplicable. Podemos pues calcular para cualquiera de las proyecciones de los ejes factoriales parámetros de autocorrelación espacial como el de Morán [4], definido por:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (8)$$

siendo

w_{ij} la distancia entre los elementos i-ésimo y j-ésimo de nuestra división territorial

y_i el valor de la proyección del elemento i-ésimo sobre las coordenadas de un factor.

$\bar{y}$ el valor medio del conjunto de $\{y_i\}_{i=1....N}$

Además podemos construir el diagrama de Moran en el que se representan los valores obtenidos frente al promedio ponderado de sus vecinos y localizar así los outliers espaciales

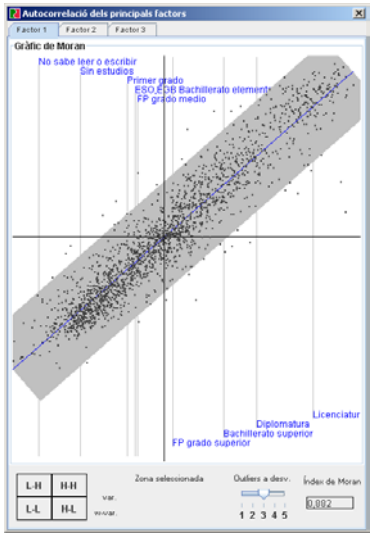

Figura 8. Gráfico de Moran. Los puntos fuera del área gris pueden considerarse outliers espaciales en relación a la proyección sobre el factor considerado del análisis de correspondencias.

4.1.3 Implementación

La implementación del AC ha sido realizado apoyándonos en el api open source JAMA[5], que implementa operaciones de álgebra matricial tales como la descomposición en valores singulares, entre otras. Nuestro plugin JUMP, modifica el esquema de las *Features* accediendo a hojas de cálculo Excel via el api open source Jakarta-POI[6], construye las matrices necesarias y realiza el análisis via JAMA. Las tablas a analizar se pueden obtener con las herramientas usuales aunque pueden también obtenerse de una forma muy cómoda mediante una aplicación basada en RMI, desarrollada en Idescat, que obtiene los agregados contra nuestra base de datos que contiene los microdatos (ya sea postGIS u otro gestor).

Una vez realizado el análisis se vuelve a modificar el esquema para añadir las proyecciones en los factores pudiendo así realizarse análisis posteriores como

clusters, utilizando para ello el api open source WEKA[7]. Los gráficos se obtienen via el api Java 3D siendo posible interactuar con ellos mediante giros, traslaciones etc.

Un punto muy importante es la interrelación de operaciones de selección entre las ventanas JUMP, los gráficos 3D y el diagrama de Moran como puede observarse en la figura 9.

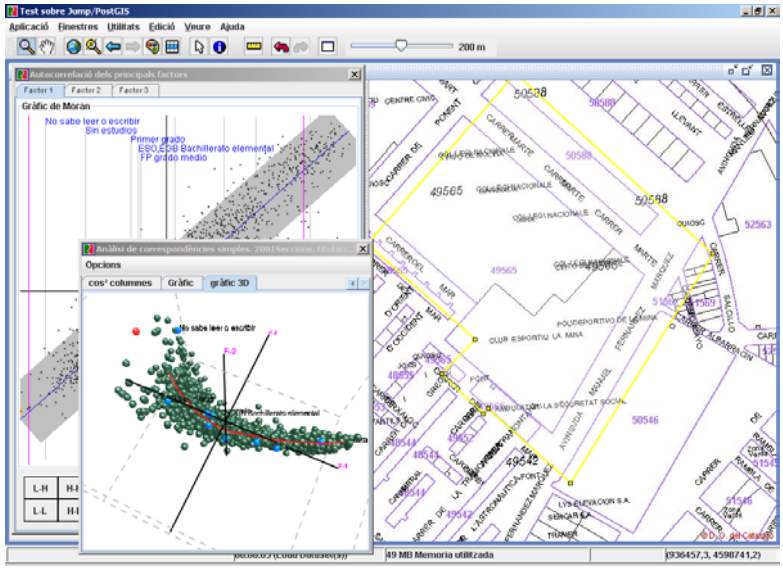

Figura 9. Tres ventanas: mapa, factorial 3D y de Moran y los efectos de una selección en el mapa: el elemento queda seleccionado en el gráfico del AC y en el de Moran.

Esto es posible gracias a que la clase JUMP *LayerViewPanel* dispone del miembro publico *addListener(LayerViewPanelListener listener)*. Indirectamente, las ventanas que incorporan el grafico 3D y de Moran implementan la interficie *LayerViewPanelListener.*

4.2. Ejemplo: población según código de actividad a 3 dígitos en el Barcelonés

El análisis realizado a la matriz de actividad vs secciones censales del Barcelonés para el censo de 2001 revela fuerte efecto Guttman en el plano F1-F3 (17.8% de inercia explicada). El factor F1 ordena los códigos de actividad oponiendo ciertos códigos a otros (figuras 11 y 12):

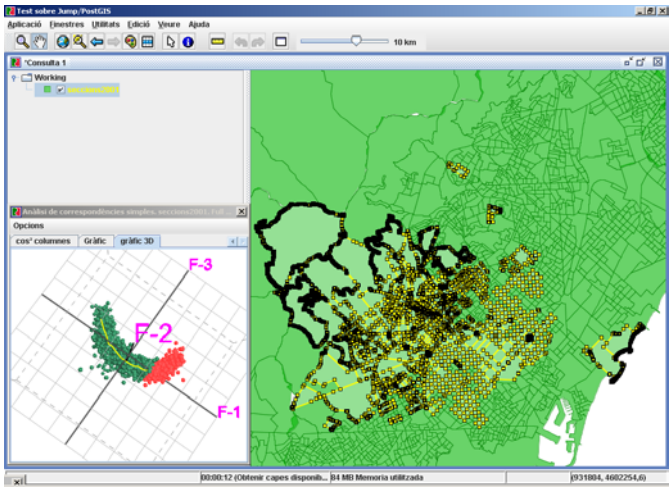

Figura 11. Valores altos en el eje F1. Elevada contribución de: actividades sanitarias, Administración Pública, actividades jurídicas y de contabilidad, Intermediación monetaria.

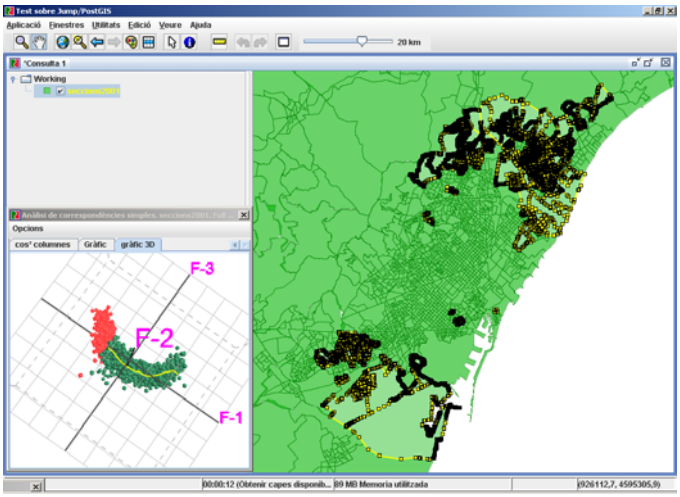

Figura 12. Valores bajos en el eje F1. Elevada contribución de: construcción general de inmuebles, personal doméstico, transporte terrestre y actividades industriales de limpieza.

El segundo factor tiene una contribución muy alta de las categorías 553 (restaurantes) y 950 (personal doméstico) y los puntos fila que se sitúan a lo largo de este eje son secciones del casco antiguo de Barcelona con una fuerte inmigración de cierto país asiático. Como puede observarse en el gráfico factorial existen una serie de puntos que podrían considerarse outliers, tal como indica la figura 13:

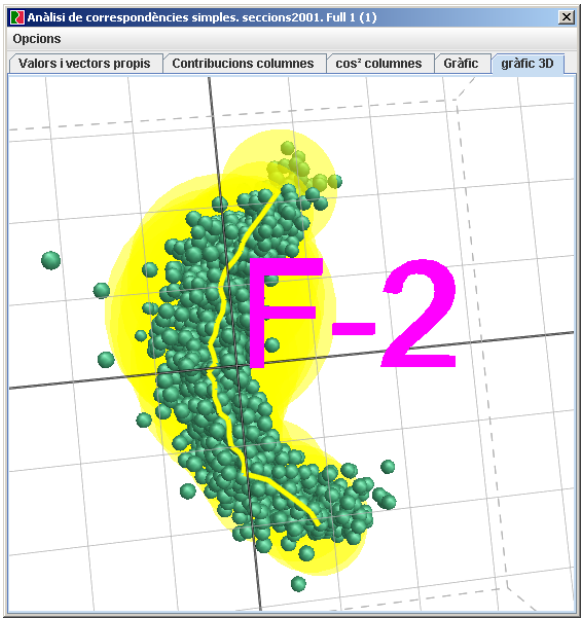

Figura 13. Outliers en la distribución. Los puntos fuera de la zona amarilla pueden considerarse outliers en la distribución.

Si el usuario selecciona uno de ellos, el elemento correspondiente queda seleccionado en el mapa, tal como queda reflejado en la figura 14:

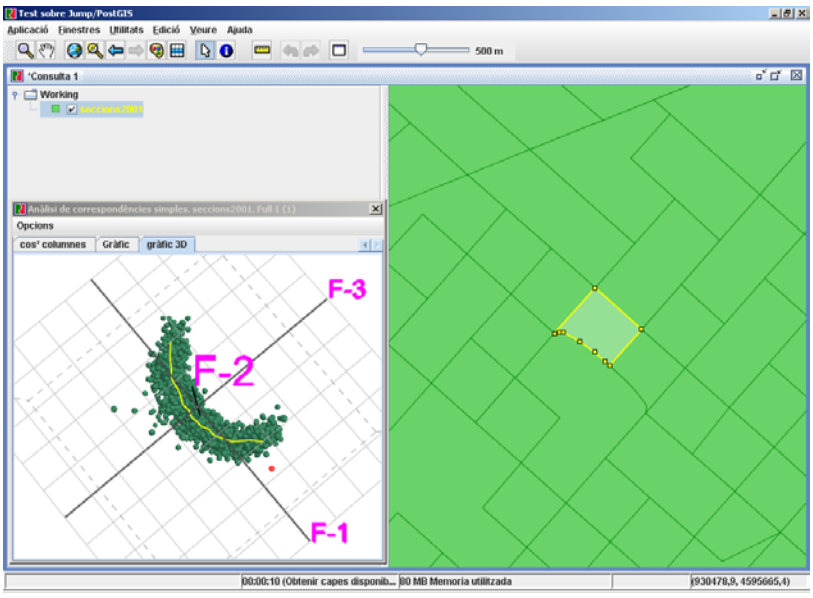

Figura 14. Un outlier seleccionado. Podría utilizase algún WMS, como el de la Dirección general del Catastro o el del Institut Cartogràfic de Catalunya para caracterizar mejor ese elemento.

El elemento seleccionado en la figura 14 se caracteriza por un elevado porcentaje de personas que trabajan en la sanidad (12.5% frente a un 5.7% para el Barcelonés). Es una sección que está al lado del Hospital Clínico de Barcelona. Otros outliers en la distribución presentan el mismo fenómeno: lo son porque cerca de ellos se encuentra o encontraba un centro especializado de trabajo produciendo una distribución

anómala de esos elementos frente al promedio. Recuérdese que la pregunta es relativa a la actividad del centro de trabajo y no debe confundirse con el de la ocupación.

El cálculo del índice de Moran sobre la proyección del primer factor (I = 0.881) indica una fuerte autocorrelación espacial. El grafico de Moran (figura 15) revela una serie de outliers espaciales, algunos de ellos ya conocidos como las secciones de la Vil·la Olímpica de Barcelona.

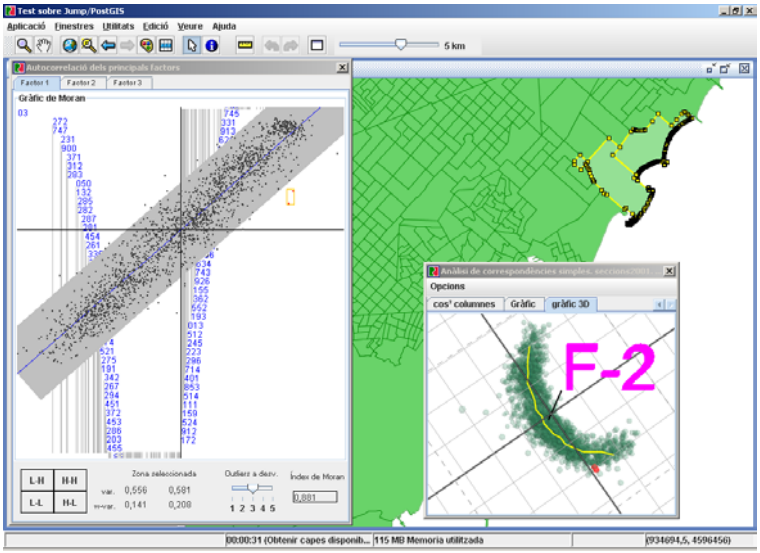

Figura 15. Unos conocidos outliers espaciales: la Vil·la Olímpica de Barcelona.

Es necesario resaltar que en el caso de la variable actividad el número de grados de libertad para el caso del Barcelonés es de 214. Aunque los tres primeros factores sólo explican un 21% del total de la inercia, el resto de factores tienen unos valores propios muy bajos y muy parecidos.

La proyección en el eje F1 ordena de alguna forma esas actividades estando esta variable, además, muy regionalizada. Para otras comarcas los resultados pueden ser muy diferentes no solo debido al comportamiento de las variables sino también al tipo de división territorial utilizada: las secciones censales. Son unas divisiones administrativas del territorio que se definen para que la población asociada sea de unos 1.000 habitantes aproximadamente.

En el Barcelonés, dada su elevada densidad de población, son detectables fenómenos de autocorrelación espacial altos para ciertas variables como la actividad económica, la ocupación, la titulación, etc. Por el contrario, en otras zonas de menor densidad de población esa esperada regionalización puede verse diluida.

5. CONCLUSIONES

La utilización de los SIG como herramientas de apoyo en las operaciones de recogida de información estadística puede ayudar a mejorar sus niveles de exhaustividad y calidad.

La georeferenciación de cuestionarios permite integrarlos en el SIG y abre diversas posibilidades de utilización de estos sistemas en las operaciones de campo. En esta comunicación hemos presentado una de las posibles.

La unión de los SIG y conocidas técnicas de análisis multivariante permite la detección de outliers, tanto en la distribución como espaciales, a diferentes niveles geográficos.

Todos estos desarrollos se han podido realizar satisfactoriamente utilizando exclusivamente software open source como JUMP, JAMA, WEKA y postGIS.

REFERENCIAS

- [1]Vivids solutions website. http://www.vividsolutions.com/
- [2]postGIS website. http://postGIS.refractions.net/
- [3]Benzécri, F.(1984): *Pratique de l'analyse des données. Analyse des correspondances & classification. Exposé élémentaire.* Paris, Dunod
- [4]Anselin, L. (1994*). Local indicators of spatial association - LISA, Techn. Rep*. 9331. Regional Research Institute, West Virginia University, Morgantown WV 26506-6825. USA
- [5]Jama :A Java Matrix Package website : . http://math.nist.gov/javanumerics/jama/
- [6]Jakarta POI website : . http://jakarta.apache.org/poi/
- [7]Weka: Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/