VI CONGRESO INTERNACIONAL DE AGRIMENSURA
2013
I CONGRESO INTERNACIONAL DE AVALUOS Y CATASTRO

GVSIG - WEKA UNA HERRAMIENTA DE MINERÍA DE DATOS INTEGRADA A UN SISTEMA DE INFORMACIÓN GEOGRÁFICA PARA EL PROCESAMIENTO DE LOS DATOS GEOGRÁFICOS CATASTRALES

Francisco D. Salas Rosette[1], Edel García Reyes[2], Yuriesky Méndez Lee[1] ,Felipe Samuel Kelly[3]

[1] GEOCUBA Pinar del Río, Cuba, fsalas@geocuba.cu, Calle Isabel Rubio # 178ª
[2] CENATAV, Cuba, *egarcia@cenatav.co.cu*
[3] GEOCUBA IC, Cuba, kelly@uct.geocuba.cu

RESUMEN

El estudio integral del medio geográfico en la actualidad constituye un reto tecnológico, debido al gran volumen de información de diversas temáticas que se acumula en las bases de datos entre ellas catastrales. El procesamiento de estos datos con el empleo de las técnicas tradicionales de análisis y del conocimiento de experto en estos dominios de datos conduce en ocasiones a la obtención de resultados basados en criterios subjetivos.

Los Sistemas de información geográfica (SIG) tradicionalmente se emplean para la recuperación, manipulación y despliegue de estos datos pero carecen de opciones que posibiliten realizar un análisis integral entre las temáticas empleando para ello nuevas técnicas de procesamiento denominadas minería de datos espaciales, que consiste en la extracción del conocimiento implícito, las relaciones espaciales u otros patrones de interés no explícitamente almacenados en bases de datos espaciales empleándose para ello diversos técnicas.

El presente trabajo aborda la integración en una plataforma SIG Libre denominada gvSIG de opciones que emplean la técnica de minería de datos denominada *Reglas de Asociación Espacial*, para determinar asociaciones implícitas en bases de datos geográficos, así como la visualización y traducción de las reglas de asociación a un lenguaje común, facilitando la realización de un análisis objetivo sobre las asociaciones espaciales almacenadas implícitamente en dichas bases de datos.

Palabras Clave: Minería de Datos, Reglas de Asociación Espacial, Catastro. WEKA, gvSIG, Postgis.

1. INTRODUCCIÓN

El extenso uso de tecnologías de sensores remotos y herramientas para la captura automática y masiva de los datos ha llevado al almacenamiento de grades volúmenes de datos espaciales en bases de datos geográficos [1]. En la actualidad constituye un reto lograr análisis integrales en grandes proyectos donde intervienen variadas temáticas ya que su procesamiento obedece a la obtención de una gran cantidad de información implícitamente contenida en los mismos. Sin embargo, la extracción y comprensión de ese conocimiento implícito es altamente deseado y plantea grandes desafíos con las tecnologías tradicionales.

Todo esto trae consigo la demanda de nuevas técnicas para la extracción de relaciones espaciales, u otros patrones no explícitos almacenados en bases de datos geográficos [2], [3], [4],[5], [6].

Se plantea que día tras día inagotablemente en todas partes del mundo se generan y almacenan cantidades inconcebibles de datos. Se estima que su volumen se duplica cada 20 meses. Es así que hoy día las organizaciones tienen gran cantidad de datos almacenados y organizados, pero no pueden sacarles provecho si no disponen de herramientas para ello.

Las técnicas tradicionales de análisis de datos no han tenido un desarrollo equivalente, pues la velocidad a la que se almacenan dichos datos es muy superior en relación con la que se analizan. Por lo tanto, existe la necesidad de generar nuevas técnicas y herramientas computacionales con la capacidad de asistir a usuarios en el análisis automático e inteligente de datos [7],[8],[9],[10], [11].

En nuestro país existen diversos programas medio ambientales entre ellos; el Programa Nacional de Cambios Globales el cual aglutina más de una treintena de temáticas, la integración de sus resultados implica movilizar un numeroso grupo de expertos, no lográndose con técnicas tradicionales de análisis establecer las asociaciones entre dichas temáticas.

Instituciones como el Ministerio de la Agricultura (MINAGRI), Ministerio de Ciencia Tecnología y Medio Ambiente (CITMA), Instituto Nacional de Recursos Hidráulicos (INRH), el Grupo Empresarial GEOCUBA, entre otras, generan un gran volumen de datos geográficos importante para la toma de decisiones.

2. OBJETIVOS

OBJETIVO GENERAL

Como objetivo General de este trabajo fue desarrollar e implementar una extensión de minería de datos espaciales, empleando la técnica de reglas de asociación para la plataforma SIG Libre gvSIG.

OBJETIVOS ESPECÍFICOS

1. Integrar una nueva opción en la plataforma SIG gvSIG para el descubrimiento de las asociaciones espaciales en bases de datos geográficos, mediante el empleo de la técnica de reglas de asociación espacial que incluya :
    - ✓ La visualización cartográfica de las reglas de asociación espacial.
    - ✓ La traducción de la regla a un lenguaje común preservando la semántica sin exigir un cumplimiento estricto de las reglas sintácticas.
2. Emplear la herramienta desarrollada para el análisis de la información catastral.

3. MATERIAL Y MÉTODO

Varios han sido los trabajos (reportes técnicos, artículos, presentaciones, etc.) que se han tomado como referencia para el desarrollo e implementación de esta extensión, el más representativo es WEKA- GDPM Integrating Classical Data Mining Toolkit to Geographic Information Systems en [12], específicamente al tratar en el procesamiento la relación entre la capa principal y las capas relevantes, por otro lado el empleo de niveles de selección y de relaciones espaciales para el análisis, hace que sea más viable y flexible almacenar y procesar los datos en Postgree/Postgis.

También fueron tomados como referencia los trabajos Application of Spatial Association Rules for Improvement of a Risk Model for Fire and Rescue Service [3] , y Application of Association Rules Discovery to Geographic Information System [13] , de los cuales se valoró la forma en que se preparaba la información así como la definición del esquema de procesamiento de la información.

Descripción de la tecnología

La extensión desarrollada se denomina extdmsar (Data Mining Spatial Association Rules) y se identifica en la carpeta de extensiones de gvSIG como: org.geocuba.extdmsar. ver Fig.1

Requerimientos Mínimos

1. gvSIG 1.9, 1.11
2. java 1.61
3. extensión org.geocuba.extdmsar
4. jai-1_1_2_01-lib-windows-i586-jre.exe

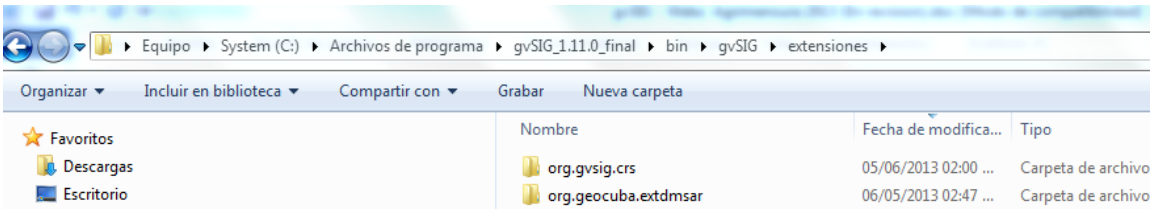

Figura No. 1 Ubicación de la extensión desarrollada

Para el trabajo con la misma cuenta con un nuevo menú denominado Reglas de Asociación ver Fig.2, este menú cuenta con cuatro opciones de trabajo, las cuales se detallan a continuación.

Figura No. 2 Menú desarrollado

Descripción de las opciones de trabajo:

La primera opción Parámetros de Conexión, nos permite establecer los parámetros para poder conectarse el usuario con la base de datos de PostgresSQL/Postgis, como se puede observar en la Fig.3

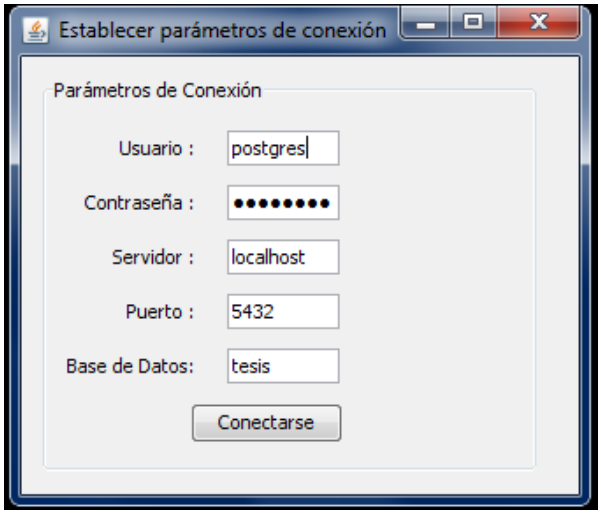

Figura No. 3 Establecer los parámetros de conexión con la base de datos

La segunda opción se denomina Taxonomía de los Datos, algunos autores hablan de taxonomía [7], otros hablan de jerarquía de conceptos [2], teniendo en cuentas que taxonomía es una jerarquía o superposición de relaciones entre diferentes categorías de un elemento, optamos por llamar de esta forma a la segunda opción del menú, siendo esta opción muy importante a la hora de definir los niveles de abstracción que deseamos para nuestro posterior procesamiento de la información.

Pudiéramos definir una taxonomía por ejemplo en la tabla parcelas como se observa en la Fig. 4, para esto establecemos varios niveles de abstracción, es importante señalar que no se deben mezclar diferentes niveles de una misma temática en la definición de la taxonomía a emplear en el procesamiento ya que esto genera información no relevante y distorsionada.

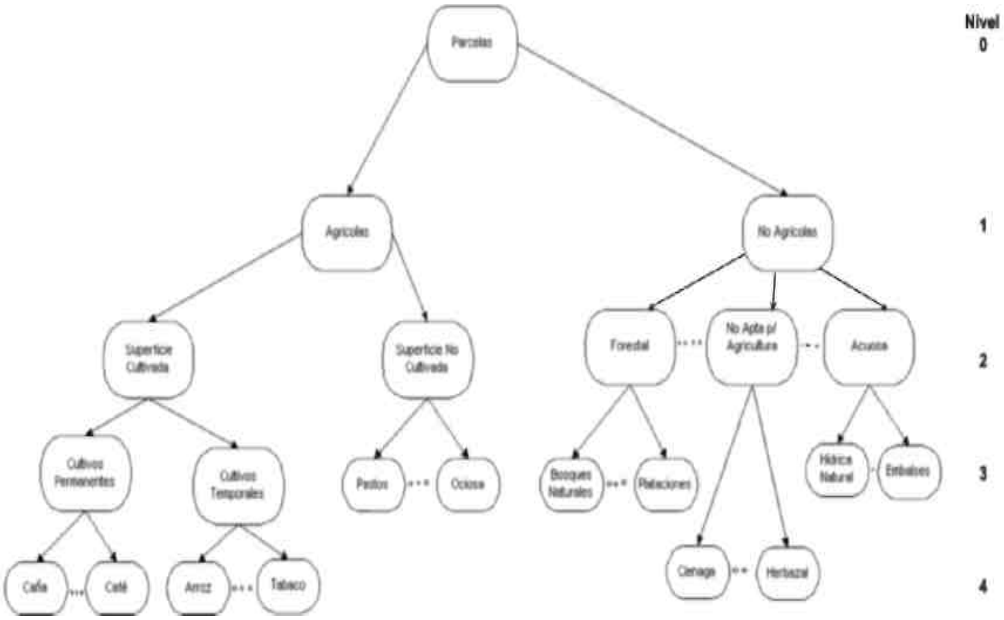

Figura No. 4 Taxonomía de parcelas

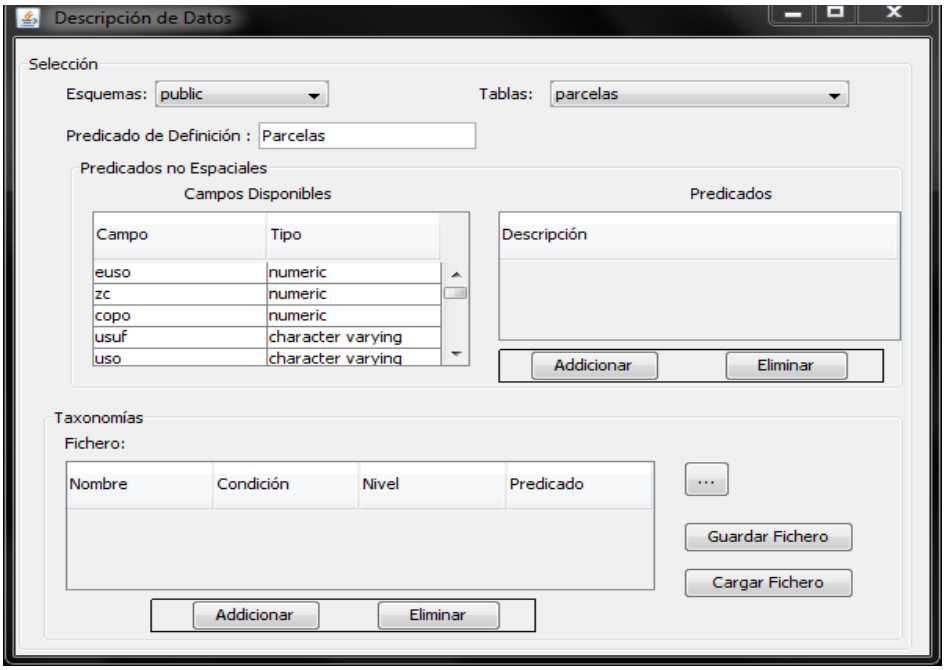

Figura No. 5 Formulario para definir las taxonomías

En el formulario de taxonomías Fig.5, es donde el usuario define los niveles de abstracción que desea procesar para sus datos, para esto se selecciona primero el esquema, luego la tabla y de forma automática se llena la tabla de campos disponibles, esto es de una gran ayuda para la hora de establecer las condiciones en la tabla de taxonomía. En la tabla de predicados, se adiciona todos los predicados que vamos a emplear, por ejemplo en el caso de parcelas uno pudiera ser Tipo_de_parcela , observe que no se debe dejar espacio en blanco para definir el nombre, y otro pudiera ser Tipo_de_Cultivo, ahora cuando comencemos a llenar la tabla de taxonomías, primeramente establecemos el nombre con que identificará (Ej, Agrícola y No_Agrícola), además sus respectivas condiciones , se debe tener en cuenta la tabla de campos (ej. Tsup = 20), después se establece el nivel, en este caso sería 1 y por último el predicado a que pertenece. Un ejemplo de nivel 2 es para el caso de los Tipos_de_Cultivos, donde el nombre Tabaco, tiene la condición tsup = 20 and tuso = 7 y nivel 2. Una vez concluido se guarda esta información mediante el botón <Guardar Fichero> en un fichero con formato xlm, teniéndose la posibilidad de cargar un fichero previamente definido y guardado.

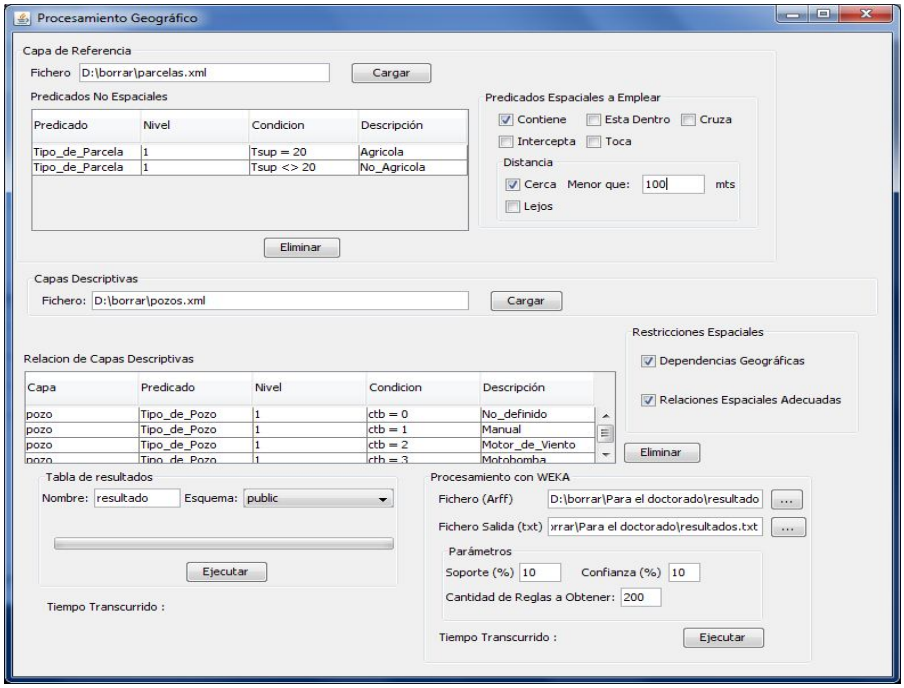

Figura No. 6 Formulario para el Procesamiento Geográfico

La tercera opción se denomina *Procesamiento Geográfico*, mediante esta opción el usuario realiza el procesamiento de la información a partir de las taxonomías creadas en la opción anterior, en este formulario Fig.6, se carga primeramente el fichero XML de la capa de referencia , llenándose la tabla de predicados no espaciales, es importante en este caso no mezclar diferentes niveles de taxonomías, para esto seleccionamos la fila y con el botón <Eliminar> , quitamos de la lista las definiciones que no deseamos procesar. Posteriormente cargamos lo referente a las capas descriptivas que deseamos procesar, cada vez que se cargue un fichero, se adicionan los datos a la tabla Relación de Capas Descriptivas, también se debe indicar que nombre tendrá la tabla de resultados la cual se almacenará en la base de datos en el esquema que se seleccione, definimos también los predicados espaciales que vamos a emplear para el procesamiento entre la Capa de referencia y las capas descriptivas, además se define el esquema y el nombre de la tabla de resultado.

Para el procesamiento de Minería de Datos, que se realiza mediante las clases establecidas en WEKA para las reglas de asociación espacial, establecemos el nombre y la ubicación del fichero arff que se genera producto al procesamiento de la tabla de resultados y el nombre del fichero de con las reglas generadas, no sin antes establecer el soporte y la confianza de la reglas.

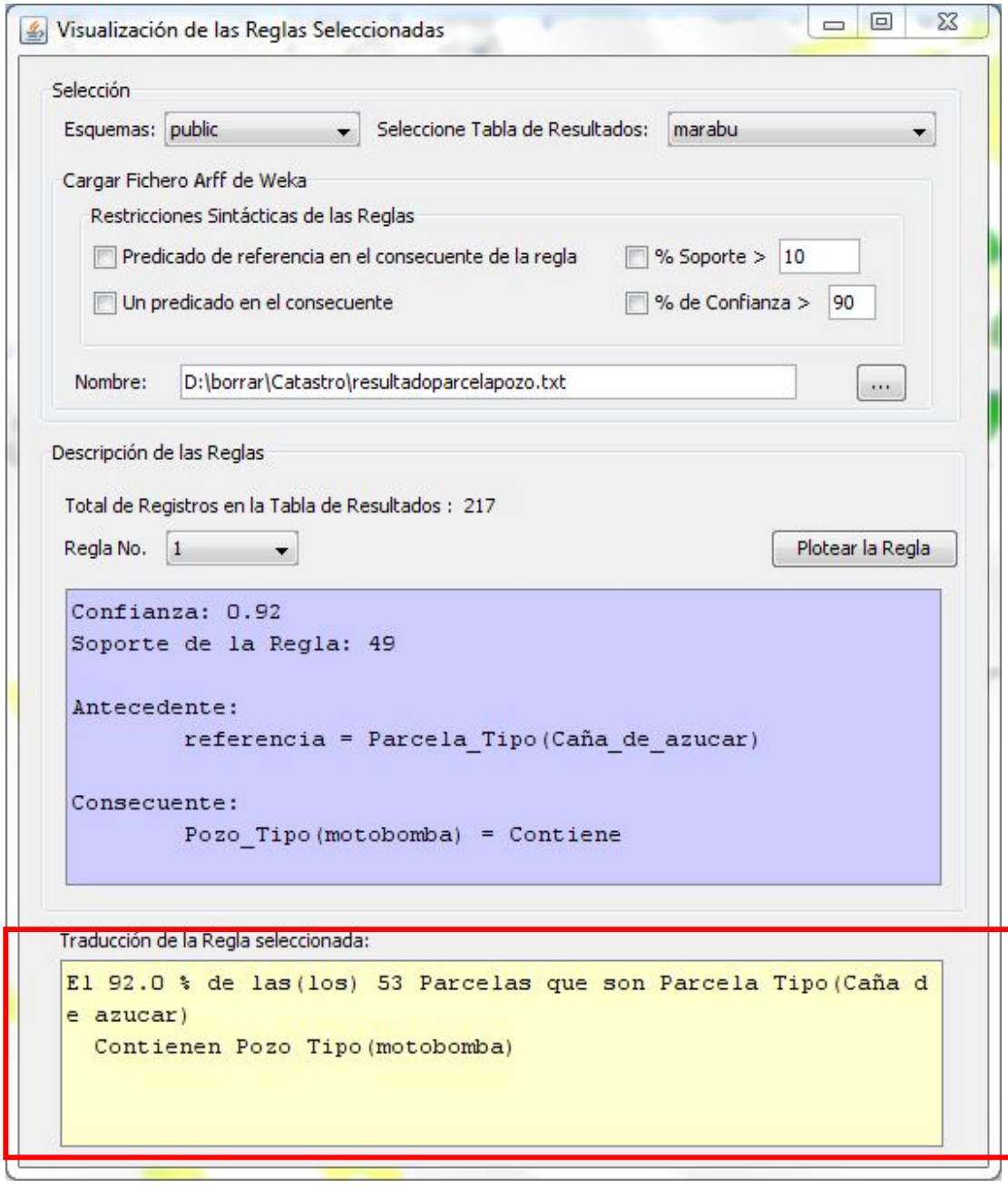

Figura No. 7 Formulario para la Visualización y Traducción de las Reglas

La opción para la Visualización de la Regla Fig.7, nos permite visualizar en el mapa la regla generada Fig. 8, así como obtener la traducción en lenguaje natural de la regla seleccionada.

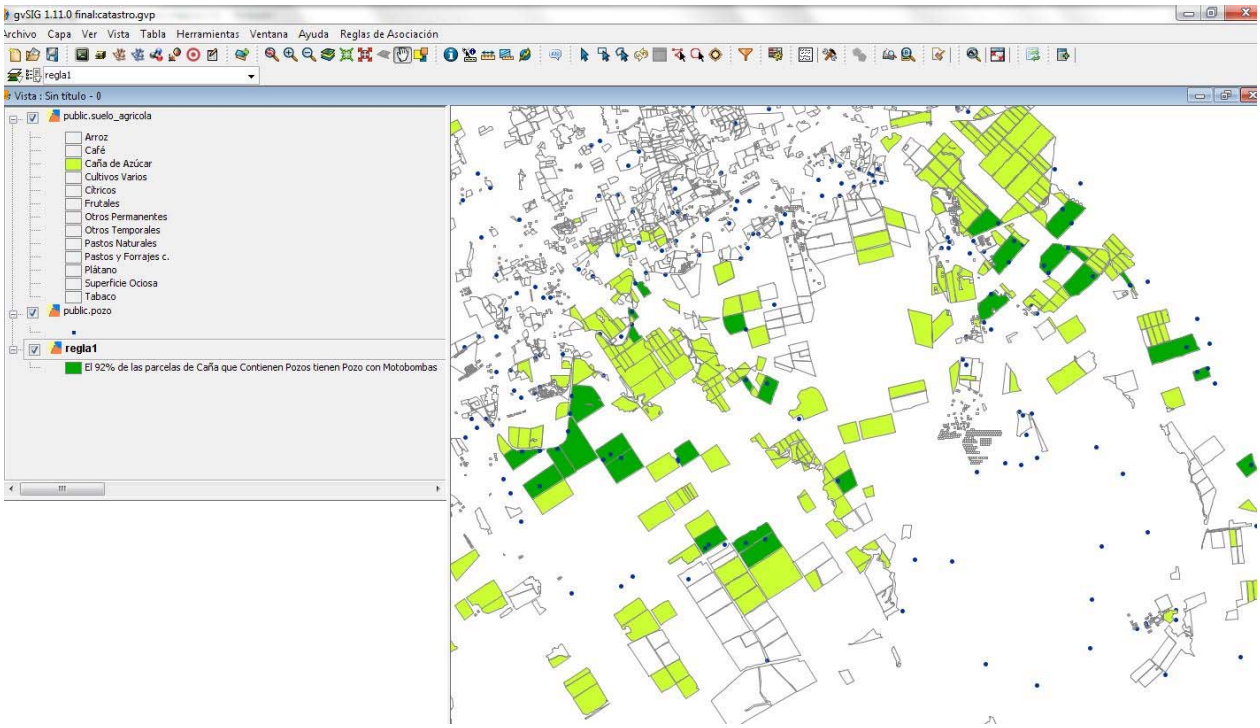

Figura No. 8 Visualización en el mapa de la regla seleccionada

RESULTADOS

1. Se logra integrar una herramienta de Minería de Datos a un SIG que posibilita visualizar en el mapa las reglas generadas y traducir a un lenguaje común dichas reglas, facilitando la comprensión de los resultados.
2. Se integra a la extensión las clases de WEKA para el procesamiento de los datos.
3. Está diseñado totalmente en software libre
4. Facilita considerablemente la comprensión del conocimiento implícito almacenado en los datos procesados.

CONCLUSIONES

La extensión desarrollada ha permitido obtener un conocimiento previamente desconocido entre las diferentes temáticas procesadas, develando en qué medida se relacionan dichas temáticas, además ha contribuido considerablemente a elevar la aplicación de estas novedosas técnicas de procesamiento de la información al lograrse una comunicación usuario - sistema de una manera más sencilla y asequible a un personal no especializado en esta materia.

REFERENCIAS BIBLIOGRAFICAS

1.S. Di Martino, et al., *Towards a flexible system for exploratory spatio-temporal data mining and visualization.* 2006: p. 12.

2. Koperski, K., et al., Discovery of Spatial Association Rules in *Geographic Information Databases.* 1995: p. 20.
3. Vera Karasova, et al., *Application of Spatial Association Rules for Improvement of a Risk Model for Fire and Rescue Services.* 2005: p. 11.
4. Karl - Heinrich Anders, O., *Data Mining for Automated GIS Data Collection.* 2001: p. 10.
5. Krzysztof, K., Jiawei Han, Junas Adhikary, *Spatial Data Mining: Progress and Challenges Survey paper.* 1996: p. 16.
6. Ding, W.a.E., Christoph F. and Yuan, Xiaojing and Wang, Jing and Nicot, Jean-Philippe, *A framework for regional association rule mining and scoping in spatial datasets.* Geoinformatica, 2011. 15: p. 1--28.
7. Kubski, M.I., *Aplicación Orientada al Descubrimiento del Conocimiento en Bases de Datos*, in *Facultad de Ciencias Exactas, Naturales y Agrimensura*2005, Universidad Nacional del Nordeste: Argentina. p. 434.
8. Jaco Pretorius, et al., *The Impact of Spatial Data on the Knowledge Discovery Process.* 2006: p. 13.
9. Martin Ester, et al., *Algorithms and Applications for Spatial Data Mining.* Geographic Data Mining and Knowledge Discovery,, 2001: p. 32.
10. Wei Ding, et al., *A Framework for Regional Association Rule Mining in Spatial Datasets.* 2006.
11. Zeitouni, K., *A Survey of Spatial Data Mining Methods Databases and Statistics Point of Views.* 1999: p. 9.
12. Vania Bogorny, et al., *Weka-GDPM – Integrating Classical Data Mining Toolkit to Geographic Information Systems.* 2006: p. 8.
13. Ansaf Salleb, et al., *An Application of Association Rules Discovery to Geographic Information Systems.* Springer-Verlag Berlin Heidelberg, 2000: p. 613-618.