PROCEDIMIENTO PARA INTEGRAR PROCESOS DE EXPLOTACIÓN DE INFORMACION Y TECNOLOGÍA GIS

Lorena E. Flores[1], Sonia I. Mariño[1], Sebastian Martins

[1]Facultad de Ciencias Exactas y Naturales y Agrimensura. Universidad Nacional del Nordeste.
9 de Julio 1449, 3400 Corrientes, Argentina.
lorenaelizabeth.flores@gmail.com, simarinio@yahoo.com

Abstract Se describe un procedimiento para la toma de decisiones integrando procesos de explotación de la información y tecnología GIS. Se plantea aplicar técnicas de minería de datos sobre la base de datos geo-espaciales para identificar patrones y comportamientos delimitado a cierto espacio geográfico. Lo expuesto permitirá potenciar los procesos de análisis de un sistema de información geográfica. Se integran dos herramientas de software libre una para construir modelos de explotación de la información y otra para la gestión de información geográfica. La validación del procedimiento se aplicó para caracterizar los robos y hurtos registrados en un semestre del año 2017 en una ciudad argentina. Los conocimientos adquiridos y productos generados son aplicables en contextos públicos y privados para la toma de decisiones basada en tecnologías emergentes.

Palabras clave: gestión de la información, diseño de procedimientos, Minería de datos, Explotación de información, GIS.

1 Introducción

La evolución y el desarrollo de las tecnologías lograron implementar innovadores métodos y herramientas para el tratamiento y análisis de la información con miras a la producción de conocimiento y apoyo a la toma de decisiones. Numerosas evidencias que ilustran múltiples aplicaciones de las tecnologías para la toma de decisiones se localizan en la literatura.

Uno de estos dominios trata el análisis del delito. Éste incluye el diseño de bases de datos espaciales, la visualización de los hechos a través de mapas y la aplicación de técnicas complejas de minería de datos (MD) [1]. Estas tecnologías de análisis incluye a los Sistemas de Información Geográfica (SIG o Geographic Information System, GIS su acrónimo en inglés), que se define como un sistema informático para capturar, almacenar, consultar, analizar y mostrar datos geo-espaciales [2].

La explotación de información (Information Mining) constituye la sub-disciplina de la Informática que aporta a la Inteligencia de Negocio [3], métodos y herramientas para la transformación de información en conocimiento [4]. Un proceso de explotación de información se puede definir como un conjunto de tareas relacionadas lógicamente [5] que se ejecutan para lograr, a partir de un conjunto de información

con un grado de valor para la organización, otro conjunto de información con mayor grado de valor que el inicial [6]

Para lograr este objetivo se utilizan las técnicas de minería de datos (data mining o DM). La minería de datos es la extracción de información no trivial, implícita, previamente desconocida y potencialmente útil de una base de datos [7]. Es un elemento fundamental para la explotación de información o bien del proceso que tiene como objetivo el descubrimiento de conocimiento en grandes bases de datos (Knowledge Discovery in Databases o KDD) [8].

Los procesos de explotación de información basados en sistemas inteligentes [9], se centran en el descubrimiento de patrones de conocimiento en la masa de datos, aplicando técnicas de minería de datos. Aun cuando la literatura es amplia se considera relevante cómo extender estas técnicas de descubrimiento de patrones en la información contenida en una base de datos geo-referenciada generada y utilizada por un GIS. También se plantea cómo visualizar en un mapa las reglas obtenidas por minería de datos utilizando herramientas de geo-referenciación con la finalidad de comprender su distribución espacial en un contexto geográfico específico.

Por lo expuesto, se establece como objetivo general diseñar un procedimiento que integre las tecnologías GIS y los métodos comprendidos en la minería de datos para aplicar procesos de explotación de la información en un dominio específico

2 Método

A continuación, se mencionan las fases que se establecieron para elaborar el procedimiento propuesto:

- Se identificaron metodologías de minería de datos.
- Se seleccionó CRISP-DM (Cross Industry Standard Process for Data Mining). Esta metodología comprende las siguientes fases de: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación e Implementación [13].
  Se estudiaron metodologías para implementar tecnología GIS.
- Se identificaron herramientas para aplicar minería de datos y tecnología GIS.
- Se estudió la forma de integrar ambas tecnologías: MD y GIS
- Se elaboró un procedimiento genérico que integra ambas tecnologías. Cabe aclarar que a efectos de validar la propuesta se optó como herramienta de MD a Tanagra y de tecnología GIS a QGIS [10, 11, 12]. Las herramientas de MD disponen de varios métodos de minería de datos de análisis exploratorio de datos, aprendizaje estadístico, aprendizaje automático y área de bases de datos. contiene algoritmos de aprendizaje supervisado, agrupamiento, análisis factorial, estadísticas paramétricas y no paramétricas, reglas de asociación, selección de características y algoritmos de construcción [10].
- Como software GIS se utilizo la herramienta QGIS (Quantum Geographic Information System) para el procesamiento de información geográfica dado que permite recopilar, almacenar, procesar, analizar, gestionar y presentar todo tipo de datos espaciales y geográficos [11]. La interfaz gráfica de usuario de QGIS es

especialmente útil para el análisis de datos, así como su digitalización y generación de mapas. La misma posee soporte totalmente desarrollado para el procesamiento de datos geo-espaciales. Además, QGIS viene con una arquitectura de complementos, para el cual los usuarios han contribuido con una variedad de extensiones, este conjunto de extensiones disponibles hace que sea fácil navegar e instalar cualquier extensión necesaria para el usuario [12].

3 Propuesta de procedimiento

El procedimiento propuesto como solución a la problemática en cuestión, describe las etapas para integrar un software de escritorio GIS a los procesos de explotación de información. Este procedimiento plantea ejecutar un proceso de minería de datos sobre la base de datos geo-espaciales para identificar patrones y comportamiento delimitado por un determinado espacio geográfico, con el fin que los procesos de análisis que proveen un sistema de información geográfica pueda ser potenciada con procesos de explotación de información [14].

Esta propuesta permite integrar una herramienta GIS dentro de la etapa CRISP-DM "Evaluación". Una vez obtenidos los datos explotados resultantes de la aplicación de los algoritmos de minería de datos, éstos se visualizan a través de un GIS para lograr un análisis más profundo con la ayuda del componente espacial. A continuación, se observa en la Fig. 1, la integración de tecnología GIS en la etapa Evaluación de la metodología de explotación de información CRISP-DM.

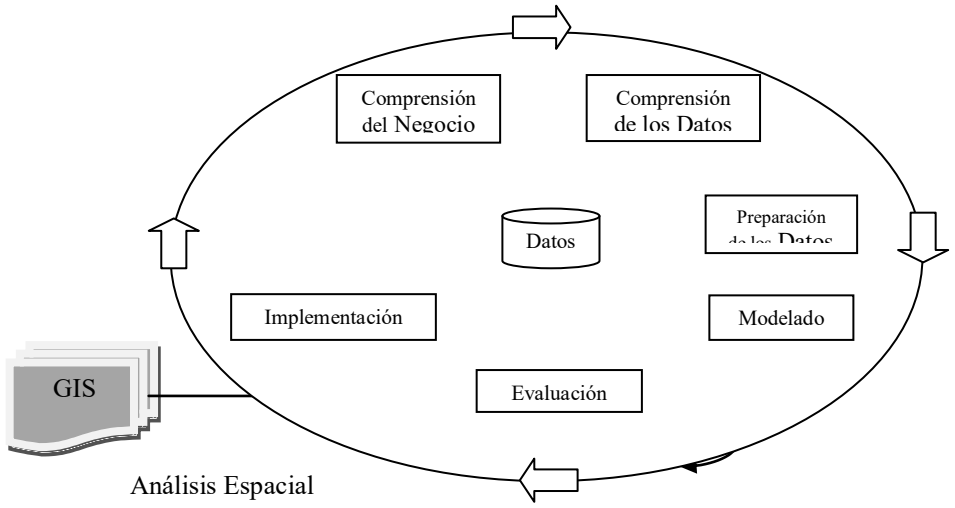

Fig. 1. Integración de tecnología GIS en CRISP-DM (Fuente propia)

3.1- Alcance del procedimiento propuesto

El alcance de este procedimiento está determinado al hallazgo de posibles patrones de comportamiento, utilizando técnicas de minería de datos, y determinadas según la problemática en cuestión. Se optará por emplear de forma conjunto una herramienta de geo-localización la cual mostrará en un mapa los resultados obtenidos de aplicar los algoritmos de minería de datos, permitiendo analizar detalladamente según zonas geográficas específicas.

3.2- Requerimientos del procedimiento propuesto

El procedimiento se integra con el acceso a los atributos de la capa vectorial geográfica incluidos en los archivos tipo shape utilizando una conversión de datos para la interpretación de la herramienta de MD. Luego se aplican las técnicas de explotación de información y el resultado se unifica con los datos espaciales a través de la unión de tablas utilizando GIS como medio.

- Se establecen como requerimientos funcionales:

El sistema permite un análisis más detallado utilizando la herramienta espacial a partir de la información proveniente de aplicar minería de datos.

El procedimiento no realiza la integración de forma automática, se debe utilizar un complemente que ofrece el software de escritorio QGIS para unir los datos.

Se precisan como requerimientos no funcionales:

- El procedimiento puede implementarse sobre plataformas Windows XP y versiones posteriores.
- El procedimiento trabajará con archivos en formato shape y debe contener los cuatro archivos básicos: shp, shx, prj y dbf.
- La integración funciona solo si se mantiene el atributo o campo de unión (campo de id del registro) en el proceso de explotación de información.
- La integración funciona sobre cualquier versión de software QGIS y de Tanagra. No se han realizado estudios de integración sobre otras herramientas.

3.3- Etapas del procedimiento propuesto. La Fig. 2 muestra la integración de tecnologías de DM y GIS . En particular las tecnologías GIS se incorporan en la fase evaluación de la metodología CRISP-DM.

Fig. 2. Etapas del procedimiento de integración MD y GIS (Fuente propia)

A continuación, se describe el procedimiento que sustenta la propuesta, identificando como herramienta de MD a Tanagra y de tecnología GIS a QGIS:

Etapa 1. Comprender el dominio del conocimiento

La primera etapa de comprensión del dominio del conocimiento hace referente al análisis del campo de estudio que se requiere analizar. Se deben tener en claro los objetivos que se desean alcanzar.

Etapa 2. Seleccionar la base de datos geográfica

En esta segunda etapa se determina la base de datos geográfica a analizar y se seleccionan los datos iniciales, se obtiene la base de datos espacial con la que se requiere trabajar y cuya capa informática debe ser de tipo vectorial, es decir, debe estar compuesta por sus cuatro archivos básicos; shp, shx, prj y dbf. Los datos geográficos se pueden visualizar a través del software GIS. Se establece como:

Inputs: shape o capa vectorial (shp, shx, prj y dbf).

Outputs: representación gráfica de capa vectorial utilizando QGIS.

Etapa 3. Convertir los datos geográficos

La etapa de conversión de datos geográficos comprende la extracción de datos de la capa vectorial (archivo con formato dbf). Extraídos los datos, se los convierte a un archivo para su interpretación por la herramienta de explotación de información (archivos con formato xls). Se establece como:

Inputs: atributos de la capa vectorial o dbf.

Outputs: archivo con formato xls.

Etapa 4. Aplicar los algoritmos de minería de datos

Efectuada la conversión de datos, en esta fase se aplican los procesos de explotación de información, es decir los algoritmos de minería de datos sobre los datos del dominio. Se recurre al archivo generado en la etapa anterior (archivo con formato xls) y se usa la herramienta de MD para aplicar los diferentes algoritmos que identifican y caracterizan los diferentes patrones de la base de datos de acuerdo a la problemática a resolver. Se establece como:

Inputs: archivo con formato xls.

Outputs: aplicación de algoritmos de minería de datos.

Etapa 5. Analizar los resultados

Esta etapa consiste en interpretar los resultados obtenidos de aplicar las técnicas de minería de datos en la etapa anterior. Se analizan y evalúan las diferentes salidas e informes obtenidos por la herramienta de explotación de información. Se establece como:

Inputs: resultados de la aplicación de algoritmos de minería de datos.

Outputs: análisis de informes (archivo con formato txt) y reportes (formato de página html) con los resultados de aplicar minería de datos

Etapa 6. Exportar el conjunto data set

Aplicados los algoritmos y el análisis de los resultados, se exporta el conjunto de datos o data set obtenidos por la herramienta de explotación de información. Se establece como:

Inputs: resultados de la aplicación de algoritmos de minería de datos.

Outputs: archivo con formato txt.

El software Tanagra dispone de la opción “Export dataset” en la pestaña general “Components” denominada “Data visualization”, la misma permite exportar los datos con los resultados obtenidos de la aplicación de cada algoritmo y cuya salida es un archivo de texto, su nombre por defecto se denomina “output.txt”.

Etapa 7. Convertir los datos explotados

Se convierte el archivo de texto exportado por la herramienta de minería de datos, por ejemplo Tanagra, a un archivo cuyo formato sea interpretado por el software QGIS, un archivo con extensión .xlsx. Se establece como:

Inputs: archivo con formato txt.

Outputs: archivo con formato xlsx.

Etapa 8. Unificar los datos

Se recurre al software GIS y se importa el archivo generado en la etapa anterior como una nueva capa vectorial (archivo con formato .xlsx), se aplica la operación “JOIN” [15], para unificar los datos resultantes de aplicar los algoritmos de minería de datos junto con la base de datos geográfica original a través de un atributo en común. Se establece como:

Inputs: archivo con formato xlsx y capa vectorial origen (shp, shx, prj y dbf).

Outputs: capa vectorial (shp, shx, prj y dbf).

Etapa 9. Representar geográficamente los datos

La unión de tablas genera una capa vectorial geo-espacial explotada que podrá representarse geográficamente por sus distintos atributos utilizando un software GIS para su interpretación definitiva. Se establece como:

Inputs: capa vectorial (shp, shx, prj y dbf).

Outputs: representación gráfica de capa vectorial con QGIS.

4 Validación del procedimiento

Los sistemas informáticos destinados a los ciudadanos se constituyen en dominio que integra métodos y aplicaciones particulares de apoyo a la toma de decisiones.

En [16] se menciona que los robos representan los hechos delictivos con mayor ocurrencia en Argentina, seguido del delito de hurto y de amenazas. En el análisis de delitos criminales, el descubrimiento de patrones significativos ha brindado la posibilidad de obtener datos de interés para interpretar y adecuar este conocimiento en la definición de los planes de prevención requeridos.

La aplicación de diversas técnicas de minería de datos sobre el campo criminal se ha convertido en una herramienta con un gran potencial que permite diseñar estrategias específicas para esta área, resultando en un proceso automático de extracción de conocimiento útil [17]. Disponer de un mapa de estos delitos tan frecuentes, la zona de ocurrencia y otras características vinculadas a los mismos introduce innovadoras modalidades de obtención de la información de apoyo a la toma de decisiones.

Construido el procedimiento, se procedió a su validación. Se optó como dominio la caracterización de robos y hurtos en una ciudad a partir de los datos registrados para el primer semestre del año 2017 en el SAT.

En este contexto, se estableció como objetivo de minería de datos: identificar y caracterizar grupos entre las zonas de mayor ocurrencia de delitos para comprender los indicadores que definan a dichas zonas. En particular se establecieron como variables los barrios y las calles. Se aplicaron como parámetros los detallados en la Tabla 1.

En el procesamiento preliminar de los datos, se descartaron los campos que no aportaron información dado que mantienen el mismo valor, y otros se transformaron en un nuevo dato con más valor según el conocimiento del experto.

Se aplicó el algoritmo Kohonen SOM, que agrupó los clusters basándose en la similitud de los valores de sus atributos. La aplicación de los algoritmos permitió determinar que los barrios con mayor cantidad de delitos son los denominados como: N° 1, N° 2, N° 3, N° 4, N° 5, N° 6, N° 7 y N° 8. Para caracterizar los grupos de las zonas con mayor ocurrencia de delitos se aplicó el algoritmo C4.5. El atributo clase se define como la variable grupo generada por el algoritmo Kohonen SOM, y los atributos de entrada seleccionados se establecen en: judisdic_policial, delito_descrip, barrio_descrip, calle, tipo_lugar, clase_arma, elemento_sustraido y tipo_ataque. La clasificación que resulta de aplicar el algoritmo C4.5 brindó 9 reglas que caracterizan a los grupos identificados según las zonas de mayor ocurrencia de delitos. Las reglas generadas permitió interpretar:

- Clúster c_som_1_1: Caracterizado por robos mayoritariamente sucedidos en los barrios N° 1, N° 2, N° 3, N° 4, N° 5, N° 6, N° 7, N° 8, N° 9, N° 10, N° 11, N° 12, N° 13 y N° 14. Representa los casos en la vía pública y registra objetos personales como el elemento más sustraído por el delincuente.
- Clúster c_som_1_2: Caracterizado por la ocurrencia de delitos en su mayoría de tipo hurto.
- Clúster c_som_2_1: Determinó delitos acontecidos en un domicilio particular y en algunos de los siguientes barrios N° 1, N° 2, N° 3, N° 4, N° 5, N° 6, N° 7, N° 8, N° 9, N° 10, N° 11, N° 12, N° 13, N° 14, N° 15, N° 16, N° 17, N° 18, N° 19, N° 20, N° 21, N° 22, N° 23, N° 24, N° 25 y N° 26.
- Clúster c_som_2_2: Particularmente los delitos ocurrieron en el interior de rodado o en un comercio, y en algunos de los siguientes barrios N° 1, N° 2, N° 3, N° 4, N° 5, N° 6, N° 7, N° 8, N° 9, N° 10, N° 11, N° 12, N° 13 y N° 14.

Tabla 1. Parámetros seleccionados para el objetivo de minería de datos definido (Fuente: elaboración propia)

Algoritmo		Justificación
Kohonen SOM		
Row Size	2	Valor por defecto
Col Size	2	Valor por defecto
Distance Normalization	Variance	
Seed Row	Standard	
C4.5		
Min Size of Leaves	5	Valor por defecto
Confidence Level	0.25	Valor por defecto
Naive Bayes		

Use laplacian probestimate	Yes	
Lambda	1	Valor por defecto

5 Conclusiones

Existen una diversidad de modelos, métodos y procedimientos que introduciendo conceptos de la Ingeniería de Software facilitan el diseño y la construcción de soluciones validables en un determinado dominio de conocimiento.

El artículo expone un procedimiento específicamente diseñado para la integración de tecnologías GIS en una etapa de un método de minería de datos ampliamente conocido como es CRISP-DM. Particularmente se introduce el uso de las tecnologías GIS en la fase "Evaluación".

Los delitos como son los robos y hurtos son uno de los problemas que mayoritariamente aquejan a los ciudadanos. Con fines de validación del procedimiento diseñado, se optó por casos delictivos de una determinada ciudad, con miras a producir información para la toma de decisiones basada en la integración de tecnologías emergentes. Otros casos relacionados con la producción de información ciudadana basada en herramientas TIC podrán validarse con el procedimiento descripto.

Referencias

[1] PEÑA SUÁREZ, A., et al. *Modelo para la caracterización del delito en la Ciudad de Bogotá, Aplicando Técnicas de Minería de Datos Espaciales*. 2017. Tesis de Maestria en Ciencias de la Información y la Comunicación. Universidad Distrital Francisco José de Caldas. Bogotá. Colombia

[2] CHANG, K. T. *Introduction to Geographic Information Systems*.7th Ed. New York: McGraw-Hill, 2014.

[3] GARCÍA-MARTÍNEZ, R.; BRITOS, P. V.; BERTONE, R.; POLLO-CATTANEO, M. F., *Towards an information mining engineering*. Software Engineering, Methods, Modeling and Teaching, 2011, p. 83-99.

[4] NEGASH, S; GRAY, P. *Business Intelligence*, In: Negash, S and Gray, P. (editores). Handbook on Decision Support Systems. 2ds. Ed. (F. Burstein and C. Holsapple, Springer), 2008, p. 175-193.

[5] CURTIS, B.; KELLNER, I. M.; OVER, J. *Process Modelling.* Communication of the ACM, 1992, 35(9): pp. 75-90

[6] FERREIRA, J. E TAKAI, O. K; PU, C. *Integration of business processes with autonomous information systems: a case study in government services*. En E-Commerce Technology, 2005. CEC 2005. Seventh IEEE International Conference on. IEEE, 2005. p. 471-474.

[7] FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. *Knowledge discovery in databases: An overview*. AI magazine, 1992, vol. 13, no 3, pp. 57-60.

[8] FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. *From data mining to knowledge discovery in databases.* AI magazine, 1996, vol. 17, no 3, p. 37.

[9] BRITOS, P.V. *Procesos de explotación de información basados en sistemas inteligentes*. 2008. Tesis Doctoral. Facultad de Informática. Universidad Nacional de la Plata.

[10] RICCO, R. TANAGRA Project. 2004 (consulta: 23 noviembre 2018). Disponible en: http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html
[11] WEGMANN, M.; LEUTNER, B.; DECH, S. (ed.). *Remote sensing and GIS for ecologists: using open source software*. Pelagic Publishing Ltd, 2016.
[12] BAGHDADI, N.; MALLET, C.; ZRIBI, M.. *QGIS and Generic Tools*. 2018.
[13] MOINE, J. M; HAEDO, A. S.; GORDILLO, S. E. *Estudio comparativo de metodologías para minería de datos.* En XIII Workshop de Investigadores en Ciencias de la Computación. 2011.
[14] FLORES, L.; MARIÑO, S.I.; MARTINS, S. *Propuesta de procedimiento para el análisis delictivo basado en la explotación de la información.* En XX Workshop de Investigadores en Ciencias de la Computación (WICC 2018, Universidad Nacional del Nordeste). 2018.
[15] ZARZOSA, N. L.; ANDRÉS, M. NÚÑEZ, A. Sistemas de información geográfica. Prácticas con Arc View. Univ. Politèc. de Catalunya, 2004.
[16] Ministerio de Seguridad de la Nación Argentina, (2019, Julio 25). "Estadísticas Criminales en la República Argentina – Año 2017 Informe". [En línea]. Disponible en: https://estadisticascriminales.minseg.gob.ar/reports/Informe%20SNIC%202017.pdf
[17] Colleen M., "Process Models for Data Mining and Analysis," in Data Mining and Predictive Analysis, 2nd ed., Butterworth-Heinemann, pp. 45-65, 2015.