Detección de Patrones de Distribución en Ecología Marina: Un Caso de Estudio

Matias Pol'la[12], Agustina Buccella [12], Alejandra Cechich [1], Maria del Socorro Doldan[3], Enrique Morsan[3] , y Maximiliano Arias [1]

1- GIISCO Research Group
Departamento de Ingeniería de Sistemas - Facultad de Informática
Universidad Nacional del Comahue - Neuquen, Argentina
{matias.polla,agustina.buccella,alejandra.cechich} @fi.uncoma.edu.ar, ariasmaxi89@gmai.com
2- Consejo Nacional de Investigaciones Cientícas y Técnicas - CONICET
3 -Instituto de Biología Marina y Pesquera "Almirante Storni"
Universidad Nacional del Comahue - Ministerio de Producción de Rio Negro
San Antonio Oeste, Argentina
{msdoldan,qmorsan}@gmail.com

Abstract

Hoy en día, el gran volumen de datos necesario para modelar las diferentes realidades y fenómenos existentes y la tendencia de crecimiento que sufren los almacenes y bases de datos produce la necesidad de contar con herramientas y métodos de minería de datos. Esto se ve reflejado en la gran cantidad de proyectos e investigaciones que se llevan a cabo para encontrar algoritmos mas eficientes para el descubrimiento de patrones. El auge de los GIS y la información geográfica también producen un gran interés en el descubrimiento de mejores técnicas que consideren diferentes aspectos inherentes a los tipos de datos utilizados y a las relaciones entre los mismos. En este trabajo se describe el análisis y las pruebas realizadas para la creación de un nuevo componente de minería de datos espacial que forma parte de una arquitectura de una SPL creada para soportar servicios dentro del dominio de ecología marina. El componente implementa una combinación de dos métodos de minería que permiten identificar patrones en la distribución de las diferentes especies pertenecientes al dominio analizado.

Palabras Clave

Sistemas de Información Geográfica, Linea de Productos de Software, Minería de Datos Espaciales, Desarrollo de Software Orientado a Componentes

Introducción

Debido al avance tecnológico y a la utilización de dispositivos tales como los GPS o herramientas de tele-detección, la información y el volumen de la misma acerca de diferentes fenómenos de naturaleza geográfica esta en continuo aumento. Esto genera que los sistemas de información geográfica (GIS, por sus siglas en inglés) sean cada vez mas populares y necesarios, y la utilización de los mismos también aumente. Por otro lado, teniendo en cuenta los nuevos tipos de datos que estos sistemas utilizan y en especial, el espacio físico requerido para poder almacenar los datos necesarios que permitan modelar los diferentes fenómenos, provoca que la búsqueda de patrones de comportamiento en la información sea una tarea poco práctica para los usuarios. Por este motivo, resulta de vital importancia la utilización de técnicas y métodos de minería de datos que posibiliten el descubrimiento de patrones o información oculta e implícita [3]. Así, la minería de datos espacial surge como una extensión a la minería de datos tradicional [5] permitiendo el análisis de tipos de datos espaciales tales como la información geo-referenciada (latitud y longitud), las operaciones topológicas (adyacencia, intersección, superposición), entre otras.

En trabajos anteriores [12,13], se describieron las actividades realizadas para el diseño y desarrollo de una línea de productos de software (SPL, por sus siglas en inglés) para sistemas de información geográficos aplicada al subdominio de ecología marina. Dicha SPL posee una plataforma de servicios comunes a todos los productos derivados de la misma, junto con un conjunto de servicios específicos que

sólo son aplicados a algunos productos. Para desarrollar la misma, se siguió una metodología para el desarrollo de SPL orientada a subdominios [13,14] implementada en una primera versión en JavaScript y luego migrada a componentes, utilizando Java. Este nuevo desarrollo permitió generar una estructura que incrementa el reuso efectivo de la plataforma y facilita tanto las modificaciones de los componentes existentes como la creación e integración de los nuevos.

En este trabajo, hemos diseñado e implementado un nuevo componente para formar parte de dicha plataforma que permite la aplicación de una combinación de métodos de minería de datos espacial para ayudar a los usuarios (biólogos marinos) de los productos desarrollados a descubrir información difícilmente identificable.

Este artículo esta organizado de la siguiente manera. A continuación se abordan los conceptos relacionados con el trabajo describiendo el marco teórico en el cual se basan las tareas realizadas en el mismo. En la Sección 3 se presentan los trabajos relacionados que combinan los sistemas geográficos con la minería de datos espacial. En la Sección 4 se muestran los avances realizados en nuestra línea de investigación en donde será implementado y verificado el componente de minería de datos espacial descripto en las secciones 5 y 6. Por ultimo, se enumeran las conclusiones obtenidas y los trabajos futuros.

2 – Marco Teórico

2.1 Sistemas de Información Geográfica

Los sistemas de información geográfica son sistemas basados en computadora diseñados para modelar, capturar, almacenar, analizar y visualizar información de naturaleza geográfica de manera eficiente [1]. Los GIS no son solo herramientas para producir mapas, son sistemas mas poderosos que integran dos tipos de datos: *datos espaciales* como los usados en la cartografía y topología en donde la base de datos constituye el mapa en sí, y *datos temáticos o no espaciales* como los usados en sistemas tradicionales como la cantidad de habitantes que tiene una ciudad, o en nuestro caso, la talla o cantidad de individuos de una especie que habitan en un determinado punto del océano. De esta manera el mapa cumple la función de proyectar una vista particular de los datos geográficos almacenados en un momento dado. Luego, con el soporte de los datos no espaciales, los sistemas geográficos proveen un número ilimitado de alternativas de análisis y de visualización para la realización de mapas desde puntos de vista distintos de acuerdo a diferentes aspectos de la información [2]. Los GIS a su vez, están compuestos por un conjunto de herramientas sofisticadas que asisten al procesamiento de la información.

En la actualidad ha habido un importante crecimiento de estos sistemas incluyendo características de los mismos dentro de sistemas específicos y también tradicionales. En los mismos, se pueden identificar tres actividades principales para el correcto uso de los datos particulares tales como son la *preparación y registro* de los datos acerca de los fenómenos de interés, el *análisis* posterior en busca de patrones o características particulares y, por ultimo, la correcta *visualización* de los datos de forma de permitirle al usuario manipularlos a través de diferentes representaciones y sacar conclusiones certeras de los fenómenos de estudio.

Han surgido también esfuerzos desde la academia e industria en estandarizar tanto la forma de representación de la información como las operaciones de soporte a la misma por parte de dos organizaciones reconocidas ampliamente en este ámbito: el Comité Técnico ISO[1] (ISO/TC 211) y el Consorcio Open Geospatial (OGC - Open Geospatial Consortium)[2]. El objetivo de ambas es estructurar y formalizar los requerimientos especiales de la información geográfica de manera de generar sistemas que maximicen

1. http://www.isotc211.org/
2. http://www.opengeospatial.org/

la interoperabilidad, modificabilidad, rendimiento, escalabilidad y reusabilidad entre otras.

Un desafío importante dentro del área de los GIS es manejar el gran volumen de datos que genera la información geográfica. El espacio necesario para almacenar cientos y miles de coordenadas por cada conjunto de objetos geográficos es un punto crítico en todo sistema. Por ejemplo la representación del perímetro de una superficie requiere miles de pares de números reales. Por este motivo los GIS deben ser soportados por bases de datos especiales que provean operaciones para un almacenamiento, consulta y acceso eficiente a la información. Esta característica produce que las transacciones para la representación, creación y modificación sean mas costosas en tiempo que para los sistemas convencionales.

Al mismo tiempo resulta muy compleja la tarea de analizar los datos en forma manual o con las herramientas de soporte de los sistemas tradicionales. Es así que existen muchos desarrollos actuales para proveer la manipulación y análisis de la información geográfica. Algunas de las más sofisticadas proveen también herramientas de minería de datos espacial [5,6,18,21] para explorar aún más este tipo de información y poder obtener un mejor análisis de la información.

2.2 Minería de Datos

La minería de datos [6] surge de la necesidad de explorar y analizar las bases de datos con grandes volúmenes de información ya que asiste al proceso de descubrimiento de patrones que permite encontrar relaciones entre los datos, o información oculta que no se presenta de manera explicita.

La minería de datos [3] se define como el *proceso por el cual se puede descubrir conocimiento interesante, implícito y previamente desconocido de grandes bases de datos*. En general en la literatura hay consenso en definir al descubrimiento de conocimiento como un proceso iterativo que involucra una serie de etapas o pasos [4]. Primeramente se requiere una etapa de *selección de los dato*s y la posterior *limpieza o depuración de los mismos* ya que es muy factible encontrarse con datos contaminados o erróneos que provoquen que los resultados finales obtenidos tras el proceso de minería sean poco precisos. Luego se debe realizar una etapa de *pre-procesamiento y transformación de los datos* en donde se asigna un formato adecuado para su posterior análisis. Para realizar los análisis se deben seleccionar diferentes algoritmos computacionales y enfoques de visualización que faciliten la interpretación y evaluación de los resultados. Es necesario repetir este proceso, reajustando los datos y variables elegidas para encontrar una mayor precisión en los resultados obtenidos. La etapa de pre-procesamiento y su transformación al formato requerido por el algoritmo a aplicar es sumamente importante ya que esto va a incidir directamente en la precisión y calidad de los resultados. Es deseable que tanto la precisión como la calidad sean lo mas altos posible ya que estos soportaran la posibilidad de descubrir patrones y conocimiento útil.

Por otra parte, la Minería de Datos Espacial (SDM, por su siglas en ingles) se define en [5], como *la extensión de la minería de datos para los sistemas de información geográfica.* Esta extensión considera todas las características inherentes a los GIS y en especial a la naturaleza de datos geográficos y espaciales con los que se trabajan. Cabe destacar que muchos de los métodos o algoritmos utilizados en la minería de datos tradicional no contemplan el hecho de que los GIS trabajen con nuevos tipos de datos, relaciones implícitas entre las variables y datos geo-referenciados. Por esta razón, la minería de datos espacial esta compuesta por varios campos, los cuales incluyen el aprendizaje automático, los sistemas de bases de datos dedicados a la información geográfica, la geometría computacional, las estadísticas y visualización de los datos, y la teoría de la información. Los métodos de

SDM pueden ser utilizados para el descubrimiento de patrones de distribución espacial así como las relaciones que existen entre datos espaciales y tradicionales.

3 - Trabajos Relacionados

Existen varias investigaciones que crean o mejoran técnicas y/o métodos para la minería de datos espacial. En [6] se describen cuatro tipos de métodos que incluyen la clasificación supervisada y no supervisada (clustering[3]), reglas de asociación espacial y geo-visualización.

Los métodos de clasificación supervisada trabajan sobre un conjunto especifico de agrupaciones de datos determinados o un grupo de diferentes categorías de acuerdo a valores de atributos previamente definidos. La aplicación de estos métodos requiere la división de los datos en dos grupos diferentes: uno de entrenamiento el cual servirá para establecer un modelo de clasificación; y uno para validar y optimizar el modelo obtenido y que sirve de evaluador del modelo ya entrenado. Dentro de este tipo de métodos se pueden encontrar las redes neuronales, los árboles de decisión, las máquinas de soporte vectorial (SVM por sus siglas en ingles) y los métodos de vecino más cercano, entre otros.

Por otro lado, los métodos basados en reglas de asociación surgen para cubrir la necesidad de hallar patrones y regularidades en los datos presentes en grandes bases de datos que tengan algún tipo de relación no explicita. Estos tipos de métodos trabajan con conjuntos de datos con los que se intenta descubrir algún tipo de comportamiento oculto formando reglas de asociación entre los elementos (campos) y las transacciones de la base de datos. El objetivo es encontrar relaciones que aparezcan con un gran porcentaje entre el conjunto de transacciones analizadas. Para las base de datos espaciales, aparecen reglas de asociación que incluyen tanto las relaciones entre los datos espaciales como cercanía, intersección, superposición y adyacencia; así como las relaciones entre estos y los datos no espaciales [7,8,9].

El tercer método que se describe en [6], intenta facilitar el descubrimiento de conocimiento a través del análisis de los datos geo-referenciados y la visualización de los mismos. La geo-visualización es definida en [10] como el método que se centra en el desarrollo de mapas interactivos y herramientas que permitan la exploración de los datos y el descubrimiento de conocimiento. A diferencia de los otros métodos que trabajan con datos tales como cadena de caracteres o dígitos, este requiere de herramientas de visualización que permitan por ejemplo el análisis de imágenes satelitales o imágenes cartográficas en formatos digitales.

Por ultimo, aparecen los métodos de clasificación no supervisada. En esta clase de métodos, a diferencia de los supervisados, no se utiliza un conjunto de entrenamiento que genere el modelo de clasificación ni de prueba para evaluarlo. Aquí el objetivo es agrupar los datos en diferentes subconjuntos significativos llamados clusters[4], en donde todos los elementos del mismo tendrán características similares y serán diferentes a los elementos pertenecientes a otros clusters. Dentro de los métodos de agrupación existe por un lado la *agrupación jerárquica*, la cual organiza los elementos en particiones anidadas, y por el otro, la *separación o partición*, en donde cada uno de los subconjuntos identificados no se solapa entre si. Los algoritmos que implementan este tipo de métodos, tales como el k-Means [11], intentan identificar a través de un número de iteraciones un conjunto de agrupaciones que son representadas por un centroide (también llamado vector media). Esta división en conjuntos se realiza al minimizar la suma de cuadrados de las distancias euclidianas entre los elementos y los centroides correspondientes.

Dentro de los métodos de agrupación para SDM [6] aparecen la agrupación espacial, la regionalización y la detección de punto caliente (hot-spot en inglés). Los primeros

3. Clustering – Término proveniente del Ingles que hace referencia a procedimientos de agrupación de una serie de vectores según diferentes criterios.
4. Cluster – Agrupación de vectores

dos son métodos bastante similares que solo difieren en que el primero trabaja con puntos aislados en el espacio, mientras que el segundo analiza zonas o regiones contiguas (grupos de objetos espaciales). El tercer método, también conocido como patrón de punto o análisis de punto caliente se centra en la detección de concentraciones de fenómenos particulares en puntos específicos. En [5] se define como el proceso de encontrar grupos de eventos inusualmente densos a través del tiempo y el espacio. Este es utilizado por ejemplo para identificar zonas de delincuencia o sectores de brotes de enfermedades.

Existen varios algoritmos que implementan métodos de agrupamiento para datos de naturaleza espacial. Entre ellos se puede encontrar PAM (Partitioning Around Medoids) [21] el cual utiliza *k* medias para identificar los distintos subconjuntos. Con el crecimiento de las base de datos y el costo de ejecución de los algoritmos surgieron diferentes variantes que intentan optimizar los tiempos, entre los cuales para SDM, se destacan CLARA (Clustering Large Applications) [21], el que particiona el conjunto de datos en diferentes muestras y aplica PAM a cada una de ellas. Otro algoritmo que es utilizado específicamente para datos espaciales es el CLARANS [17] (Clustering Large Applications based upon Randomized Search) que es una fusión de los dos anteriores. En [18] se presentan dos variantes de CLARANS con diferentes enfoques. El primero SD (Spatial Dominant) esta enfocado en los datos espaciales y su objetivo es el descubrimiento de patrones de datos no espaciales en agrupaciones o subconjuntos que presentan una relación espacial. El segundo enfoque, NSD (non-spatial Donminant), realiza el trabajo inverso, intentando identificar agrupaciones de datos espaciales a través de un conjunto de datos de naturaleza no espacial. Estos, primero aplican alguna técnica de generalización, la cual realiza una clasificación de acuerdo a una jerarquía de conceptos previamente definida, para luego aplicar agrupación (clustering) considerando la información espacial.

Otro algoritmo de agrupamiento es el DBSCAN [22], que puede ser utilizado para descubrir diferentes agrupaciones de forma arbitraria y también puede ser aplicable a datos espaciales. Este presenta la ventaja de identificar ruido entre los datos procesados ya que su funcionamiento produce la identificación de agrupaciones densas, relegando instancias aisladas. Además, otra ventaja es que presenta un menor tiempo de procesamiento en comparación con CLARANS[22], lo que provoca su elección sobre este ultimo para grandes conjuntos de datos.

Por ultimo en [11] se realiza un análisis de minería de datos espacial utilizando la herramienta de aprendizaje WEKA[5] y aplicando el método de agrupamiento no supervisado k-Means. WEKA tiene la ventaja de presentar una sencilla interfaz y esta desarrollado en Java lo que permite ser utilizado y ensamblado en cualquier aplicación. Además incluye una gran variedad de algoritmos de minería de datos que vienen incluidos y que permiten la realización de tareas tales como la clasificación, regresión y clustering, entre otras. Esto, sumado a la posibilidad de incorporar nuevas funciones de manera sencilla, lo posiciona entre los mas usados.

Sin embargo, a pesar de existir mucho trabajo de investigación con respecto a las herramientas y algoritmos de minería de datos espaciales, se ven pocas aplicaciones reales en la literatura. La mayoría de las propuestas intentan mejorar los tiempos de ejecución olvidando evaluar si los resultados obtenidos son los correctos o necesarios. Así, en este trabajo se ha evaluado la utilización de algunos métodos de minería de datos espacial centrándonos en los resultados obtenidos y su utilidad. Así surgió una combinación de algunos ya propuestos, permitiéndonos brindar una nueva funcionalidad a sistemas geográficos que asisten a requerimientos especiales de usuarios de los mismos en el dominio de ecología marina.

5. WEKA - http://www.cs.waikato.ac.nz/ml/weka/

4 - Antecedentes

En trabajos previos [12,13,14] hemos descripto la investigación realizada para incrementar el reuso efectivo en el domino geográfico. Para esto se ha diseñado e implementado una Línea de Productos de Software (SPL por sus siglas en inglés) [15,16] orientada a subdominios geográficos [12,13].

La creación de la línea se ha realizado en conjunto con dos organizaciones (IBMPAS[6] y CENPAT-CONICET[7]) que trabajan en el subdominio de Ecología Marina y que nos han proporcionado los requerimientos necesarios para llevar a cabo sus actividades diarias.

El dominio de Ecología Marina se centra en el estudio de las relaciones de todos los organismos que viven en el hábitat de la vida marina y la interacción de estos con su ambiente. Involucra los *factores abióticos*, que incluyen todo que aquello que es inerte y no presenta vida como por ejemplo la temperatura, productos químicos, la salinidad, la luz, la profundidad; y *los factores bióticos*, que incluyen todos los organismos vivos (flora y fauna) que habitan en un ecosistema. Los estudios en el ámbito de la ecología marina abarcan el examen de los microorganismos unicelulares, el entorno donde habitan las especies, el impacto de la actividad humana y los efectos globales de la contaminación.

La SPL creada para este subdominio incluye una serie de funcionalidades o servicios que son útiles para realizar el conjunto de actividades necesarias y arribar así a conclusiones acerca de la vida y la conservación de los organismos.

En particular, la información se recopila a través de censos o campañas donde se realizan muestreos sobre las especies que viven en el mar. En nuestro caso, se realizan diferentes análisis sobre especies que habitan en los golfos San Matías, Nuevo y San Jorge (Patagonia-Argentina). Cada golfo esta dividido en varias zonas y los censos realizados están asociados a las mismas; ya que es de interés para los biólogos marinos analizar la distribución de diferentes especies en cada una de estas zonas. Por ejemplo, en el caso del IBMPAS se llevan estudios acerca de la distribución del bivalvo "*Viera Tehuelche*" mientras que en el CENPAT-CONICET es de interés analizar el avance del alga invasora "*Undaria Pinnatífida*". Como puede observarse en la Figura 1 el modelo de datos que presenta la SPL contempla tanto los datos no espaciales como los espaciales. Dentro de los datos no espaciales podemos encontrar toda aquella información referente al dominio de la ecología marina como por ejemplo lo son las especies, su talla y su biomasa. Por otro lado, puede observarse los datos geo-referenciados propios de los sistemas de información geográfica tales como la longitud y latitud que posiciona en el espacio a cada una de las estaciones.

La SPL creada contempla los servicios necesarios para llevar a cabo el conjunto de actividades descripto previamente. La misma esta dividida en dos grandes conjuntos de funcionalidades. Por un lado tenemos la plataforma de servicios que representa el conjunto de funcionalidades comunes de todos los productos resultantes de la línea junto con las variabilidades que permiten realizar ciertas configuraciones particulares para cada uno de ellos. De esta manera, la plataforma es lo suficientemente flexible para adaptarse a diferentes requerimientos de diferentes productos. Por el otro lado, cada producto derivado de la línea esta conformado por un conjunto de funcionalidades específicas y particulares.

Para poder soportar el desarrollo de nuevos productos de la SPL fue necesario aplicar otra tecnología orientada al reuso de software, como lo es la Ingeniería de Software Basada en Componentes [19,20]. La misma ofrece herramientas que permiten el uso y ensamblaje de piezas desarrolladas por separado provocando tanto la reducción de costos como del tiempo de desarrollo.

6. IMBPAS – Instituto Biología Marina y Pesquera Almirante Storni - http://www.ibmpas.org/
7. CENPAT-CONICET – Centro Nacional Patagónico - http://www.cenpat.edu.ar/

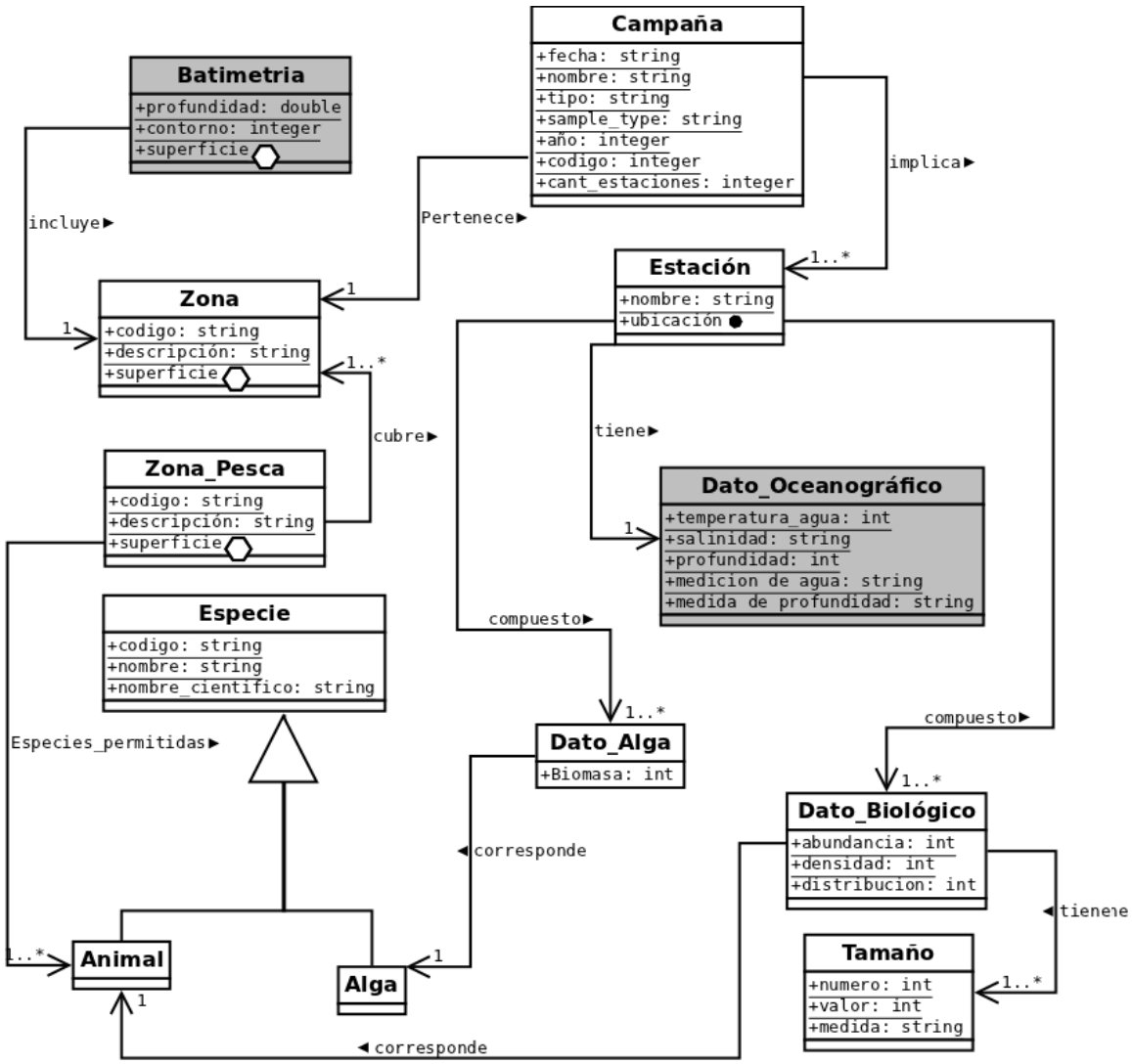

FIGURA 1 -Modelo conceptual para el domino de Ecología Marina

En la Figura 2, puede observarse la arquitectura final de la SPL, la cual esta formada por 3 capas entre las cuales podemos observar la *Interfaz de Usuario*, *Procesamiento Geográfico* y *Modelo Geográfico*; y en donde cada una de ellas esta conformada por un grupo de componentes. En la misma se pueden observar los componentes de acuerdo al dominio al que pertenecen, ya sea el dominio geográfico, el oceanográfico o el subdominio de la ecología marina. De esta forma, los componentes mostrados en gris oscuro implementan servicios que pertenecen al dominio geográfico genérico. Por ejemplo, el componente *Características de Visualización* está diseñado para implementar servicios tales como refresco, zoom, escalas, etc. Los componentes en un gris suave implementan servicios del dominio oceanográfico como por ejemplo el componente *Visor de Capas de Oceanografía* que se encarga de la visualización de algunas características de los océanos como la salinidad o profundidad. Por último, los componentes en blanco implementan servicios específicos del dominio de ecología marina. Por ejemplo, el componente V*isor de Capas de Ecología Marina* y *Visor de Atributos de Ecología Marina* permiten la visualización de las capas y atributos de la información perteneciente a las especies y zonas de interés. Por otro lado, los componentes representados por líneas punteadas indican componentes externos,

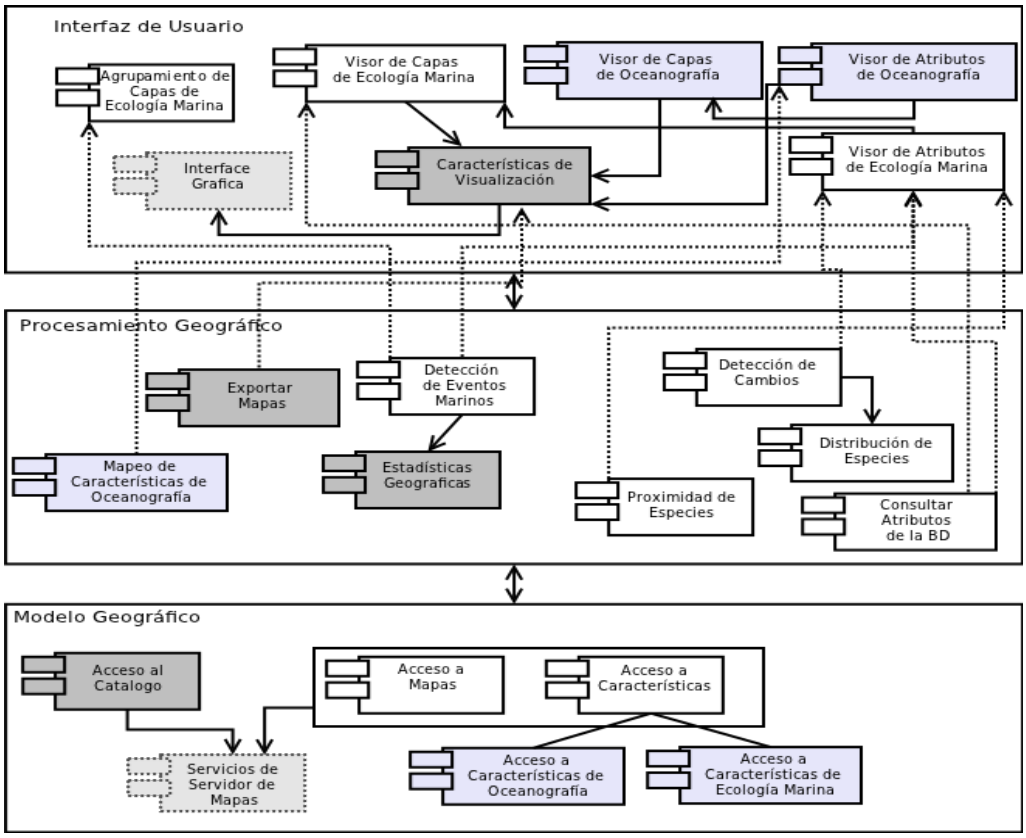

FIGURA 2 - Arquitectura de referencia de la línea de productos en el domino de Ecología Marina

como GeoServer (servidor de mapas), utilizados para implementar nuestros servicios.

De la SPL han surgido dos productos de software que permiten realizar una serie de actividades a los usuarios de los mismos. Uno de ellos, el sistema implementado para el IBMPAS, permite entre otras cosas, la visualización de las campañas, estaciones y zonas, así como la utilización de servicios de creación de histogramas para observar diferentes estadísticas de los diferentes fenómenos de estudio. El mismo esta disponible en http://gissrv.fi.uncoma.edu.ar /SaoProjectUI. El otro, implementado para el CENPAT-CONICET permite además la creación de diferentes mapas para visualizar dos o más zonas en paralelo. También presenta un componente de procesamiento encargado de analizar el avance (crecimiento y distribución) de determinadas especies a lo largo de un determinado lapso. El mismo esta todavía en etapa de validación y se encontrará disponible en breve.

La SPL y la derivación de los productos previamente mencionados, fue descripta en [14] en donde además se efectuó una validación empírica de los beneficios derivados del reuso de la aplicación e instanciación SPL.

A su vez, a pesar de que los resultados fueron alentadores, en esta primera evaluación, se analizó la posibilidad de continuar ampliando la oferta de servicios más sofisticados que permitan a los usuarios (biólogos marinos, en este caso) descubrir información útil para sus trabajos. De esta manera, se continuó trabajando en la creación de componentes que permitan la utilización de metodologías de minería de datos espaciales. La siguiente sección describe el trabajo de investigación realizado para tal fin.

5 – Componente *"Detección de Patrones de Distribución"* para Minería de Datos Espaciales

Considerando las características que presenta la SPL desarrollada anteriormente

y analizando el dominio de aplicación de la misma, se presentó un marco ideal para llevar adelante el desarrollo de un componente que se encargue de realizar minería en búsqueda de posibles patrones en los datos almacenados. De esta manera se permite el hallazgo de información interesante para las diferentes investigaciones que llevan adelante los biólogos de las organizaciones dentro del dominio de ecología marina. Este tipo de funcionalidad se consideró de interés para todos aquellos productos derivados de la SPL, por lo que este componente fue desarrollado para formar parte de los servicios que componen a la plataforma.

Para llevar a cabo el proceso de minería, se utilizó como soporte la herramienta GNU llamada WEKA la cual podía ser fácilmente integrada como un nuevo componente a la arquitectura de la SPL posibilitando y facilitando la comunicación con otros componentes del sistema. De esta manera, nuestra arquitectura, implementada específicamente con EJB[8], y diseñada para permitir la fácil evolución del sistema, posibilitó que la tarea de crear un nuevo componente que interactúe con los demás desarrollados (Figura 2) sea relativamente sencilla. Este nuevo componente, denominado *Detección de Patrones de Distribución*, permite realizar diferentes análisis de acuerdo a las distintas especies que se estudian en el subdominio de ecología marina y en particular las estudiadas por las dos organizaciones involucradas (IBMPAS y CENPAT-CONICET).

El componente permite la aplicación de una combinación de métodos de minería de datos espaciales para la identificación de patrones de distribución de especies de acuerdo a las campañas realizadas en las diferentes zonas de estudio. Para ello se tuvieron en cuenta diferentes variables tales como longitud y latitud de cada una de las estaciones que conforman las campañas dentro de las diferentes zonas, la profundidad de cada una de ellas, la cantidad de individuos de una especie en particular y la talla promedio registrada. De esta forma, las agrupaciones no solo se rigen por su posicionamiento geográfico (datos espaciales), sino que también influyen datos no espaciales que permiten obtener conclusiones más certeras acerca de los conjuntos identificados.

En la Figura 3 se puede observar los componentes que se comunican con el nuevo componente *Detección de Patrones de Distribución* para llevar a cabo el servicio de análisis de datos. Para lograr el objetivo participan los componentes de *Consulta de atributos de la BD y WEKA* de la capa de procesamiento geográfico. Dentro del componente creado se formatean los datos (archivo .arrf) y se selecciona y configura el algoritmo de minería.

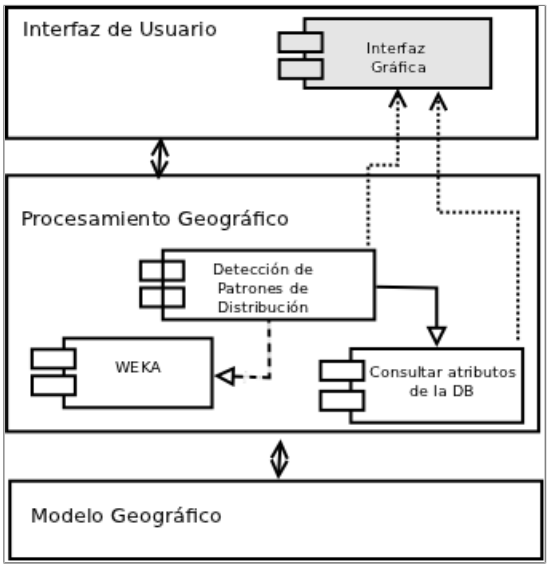

FIGURA 3 – Componente de Minería de Datos y relaciones con los demás componentes dentro de la arquitectura de la SPL

El componente utiliza como base el método de agrupamiento (clustering) de datos, ya que no se contaba con una hipótesis que permita el agrupamiento previo o la definición de un modelo de clasificación. Luego, combina métodos de minería de datos espacial de manera de obtener resultados que sean de utilidad para los usuarios del dominio de Ecología Marina. Dichos resultados fueron evaluados empíricamente por los diferentes usuarios en el conjunto de pruebas efectuadas.

8. Enterprise Java Beans - http://www.oracle.com/technetwork/java/javaee/ejb/index.html

En la siguiente sección se describen algunas de estas pruebas realizadas junto con la aplicación del mismo a uno de los sistemas derivados de la SPL.

6- Caso de Estudio: Aplicación del Componente *"Detección de Patrones de Distribución"*

Para las primeras pruebas se utilizó el sistema creado para el IBMPAS en donde fue seleccionada la especie "*Viera Tehuelche*" perteneciente a los registros de las diferentes estaciones de las campañas realizadas en la zona denominada "Playa Orengo" del Golfo San Matías, en San Antonio Oeste, Argentina. Para llevar a cabo el análisis de distribución, se realizaron pruebas aplicando los algoritmos *Simple K-Means y DBSCAN*.

En un experimento preliminar se aplicaron los métodos de forma independiente obteniendo diferentes resultados. Por ejemplo, al momento de aplicar el algoritmo simple K- Means, se identificaron los diferentes clusters sin tener en cuenta los datos erróneos (ruido) como por ejemplo, los puntos fuera de zona blanca Figura 4. Es decir, los resultados obtenidos eran poco precisos ya que los centroides de las diferentes agrupaciones estaban influenciados por el ruido presente en los datos. Por otro lado se realizaron experimentos con el algoritmo DBSCAN, cuya aplicación identificaba aquellos puntos geo-referenciados que contenían datos con ruido, ya que dicho algoritmo trabaja identificando grupos densos. A pesar de resolver la problemática de los datos erróneos, con este algoritmo no obteníamos el resultado esperado ya que solo identificaba un cluster, compuesto por la totalidad de los datos sin ruido, por ejemplo, los puntos dentro de la zona blanca Figura 4.

Por este motivo se decidió realizar una combinación de los métodos tomando las ventajas que presenta cada uno y así solucionar ambas problemáticas encontradas. Así, una vez que el usuario envía la solicitud con la zona y especie seleccionada, se aplica DBSCAN solo a los datos geo-referenciados (latitud y longitud); de esta forma se eliminan las instancias que presentan ruido. Luego se incluyen los datos no espaciales (profundidad, cantidad de individuos y talla) y se aplica Simple-K-Means configurando el número de agrupaciones que se desea identificar. Esta simple modificación y combinación de dichos métodos produce resultados muy diferentes y a la vez útiles para los usuarios.

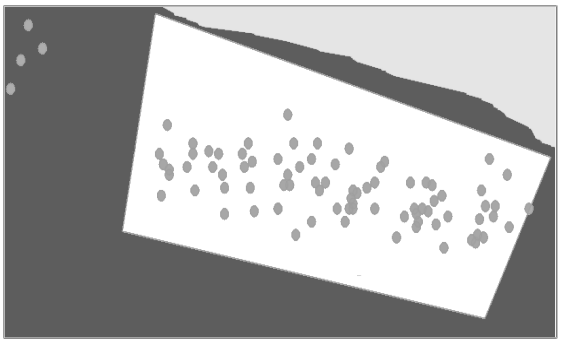

FIGURA 4 – Estaciones erróneas Zona "Playa Orengo" - Golfo San Matías – Patagonia Argentina

En las Figuras 5 y 6 se muestran los resultados obtenidos aplicando esta combinación. La misma fue ejecutada de manera de identificar 2 y 3 centroides, es decir, con el parámetro *k=2 y 3* respectivamente. A priori, los resultados obtenidos en los análisis mostraban una distribución uniforme ya que los clusters encontrados presentaban características similares en cuanto al número de instancias. Por ejemplo, en el experimento con *k=2* los centroides identificados registraron porcentajes similares de concentración de estaciones, lo que sugiere una distribución suficientemente pareja de la especie a lo largo de la superficie de la zona. Sin embargo, al aumentar el numero de clusters (*k=3 y 4*) se observo que el algoritmo identificaba un cluster con mayor porcentaje de instancias. Esto generó el análisis de los datos recolectados en las diferentes estaciones de esta campaña y se descubrió que existe una mayor concentración de individuos de *"Viera Tehuelche"* al este de la zona de estudio

como puede observarse en la Figura 6 (*k =3*).

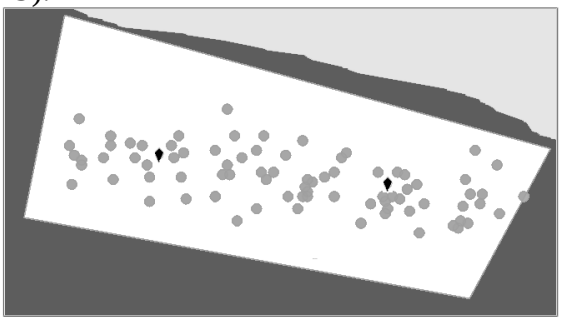

FIGURA 5 – Resultados aplicando el algoritmo Simple K Means con k = 2

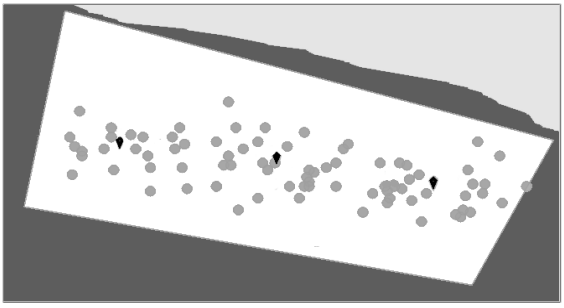

FIGURA 6 – Resultados aplicando el algoritmo Simple K Means con k = 3

Estas mismas pruebas fueron realizadas con otros conjuntos de datos (para otras especies en otras zonas) obteniendo resultados similares.

De esta forma, este nuevo componente resulta de gran utilidad para las organizaciones de biología marina ya que les permite poseer un conocimiento previo sobre el tamaño y distribución de las especies ubicadas en el mar y les facilitará las tareas en las futuras campañas. Se pueden llevar a cabo estudios en zonas más específicas, lo que generaría menores gastos tanto en el movimiento de equipamiento como en los tiempos de las campañas, al prever por ejemplo el punto espacial donde se encuentran más especímenes de una determinada especie.

Conclusiones y Trabajos Futuros

En este trabajo, se analizaron diferentes métodos de minería de datos espaciales de manera de descubrir patrones en los datos de utilidad para los usuarios del sistema. Así, por ejemplo, con la información aportada por el IBMPAS y el análisis desarrollado se puede localizar una zona más abundante en cuanto a determinadas especies de estudio. Esto genera reducción en costos por campaña y la posibilidad de un mayor numero de individuos recolectados.

A su vez, el diseño y desarrollo del componente *"Detección de Patrones de Distribución"* descripto en este trabajo, mostró un aspecto beneficioso de la SPL, su capacidad evolutiva. El nuevo componente fue fácilmente adaptado a la plataforma existente para formar parte de la misma en los productos existentes y futuros.

Como trabajo futuro es necesario realizar diferentes análisis para validar el componente implementado con una mayor cantidad de datos considerando campañas de diferentes años y épocas del año, ya que existe la posibilidad de obtener patrones erróneos de distribución.

Por otro lado, se esta trabajando en la inclusión y/o combinación de nuevos métodos de minería para datos espaciales para obtener mejores resultados y mayores precisiones. Por ultimo, considerando el componente desarrollado dentro de la SPL, es de utilidad ofrecer a los productos de la linea la posibilidad de configurarlo a través de alguna variabilidad que permita seleccionar la búsqueda de diferentes patrones de datos de acuerdo a los intereses de los usuarios.

Referencias

1 - M.A. Rodríguez Luaces. A Generic Architecture for Geographic Information Systems. PhD thesis, Univerdade da Coruña, 2004.

2 - P. Burrough and R. McDonnell. Principles of Geographical Information Systems. Oxford University Press, 1998.

3 - Frawley, W., G. Piatetsky-Shapiro, C. Matheus (1991): Knowledge discovery in databases: An overview, in G. Piatetsky-Shapiro & W. Frawley, eds, 'Knowledge Discovery in Databases', AAAI/MIT Press, Menlo Park, CA, pp. 1-27.
4 - Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery—an review. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusay (Eds.), Advances in knowledge discovery (pp. 1–33). Cambridge, MA: AAAI Press/The MIT Press .
5 - Karl-Heinrich Anders, Oberkochen (2001) Data Mining for Automated GIS Data Collection. D. Fritsch & R. Spiller, Eds.
6 - Diansheng Guo, Jeremy Mennis b (2009) Spatial data mining and geographic knowledge discovery—An introduction. Computers, Environment and Urban Systems 33, 403–408
7- Appice, A., Ceci, M., Lanza, A., Lisi, F. A., & Malerba, D. (2003). Discovery of spatial association rules in geo-referenced census data: A relational mining approach. Intelligent Data Analysis, 7(6), 541–566.
8 - Mennis, J., & Liu, J. W. (2005). Mining association rules in spatio-temporal data: An analysis of urban socioeconomic and land cover change. Transactions in GIS, 9(1), 5–17.
9 - Han, J., Kamber, M., & Tung, A. K. H. (2001). Spatial clustering methods in data mining: A survey. In H. J. Miller & J. Han (Eds.), Geographic data mining and knowledge discovery (pp. 33–50). London and New York: Taylor and Francis.
10 - MacEachren, A., & Kraak, M.-J. (2001). Research challenges in geovisualization. Cartography and Geographic Information Science, 283–312.
11 - R. Sharma, M. Afshar Alam & A. Ran (2012) K-Means Clustering in Spatial Data Mining using Weka Interface. International Conference on Advances in Communication and Computing Technologies (ICACACT) . 26 - 30
12 - P. Pernich, A. Buccella, A. Cechich, M. S. Doldan, E. Morsan, M. Arias, and M. Pol'la. (2011) Developing a Subdomain-Oriented Software Product Line. CACIC 2011: XVII Congreso Argentino en Ciencias de la Computación.
13 - P. Pernich, A. Buccella, A. Cechich, M. Pol'la, M. Arias, M. S. Doldan and E. Morsan. (2012) Product-Line Instantiation Guided By Subdomain Characterization: A Case Study. Journal of Computer Science and Technology 2012, Special Issue 12(3). ISSN: 1666-6038. Disponible en http://journal.info.unlp.edu.ar/journal/journal34/this.html **(116-122).**
14 – A. Buccella and A. Cechich and M. Pol'la and S. Doldan and E. Morsan (2013) Towards Systematic Software Reuse of GIS: Insights from a Case Study. Computers & Geosciences. 54. Elsevier Science Publishers B. V. ISSN: 0098-3004, DOI: 10.1016/j.cageo.2012.11.014. (9-20).
15 - Klaus Pohl, Günter Böckle, and Frank J. van der Linden. (2005) Software Product Line Engineering: Foundations, Principles and Techniques. Springer-Verlag New York, Inc., Secaucus, NJ, USA.
16 - Frank van der Linden, Klaus Schmid, and Eelco Rommes. (2007) Software Product Lines in Action: The Best Industrial Practice in Product Line Engineering. Springer-Verlag New York, Inc.,Secaucus, NJ, USA.
17 - Raymond T. & Jiawei Han .CLARANS-A Method for Clustering Objects for Spatial Data Mining.
18 - Kaufman, L., and P. J. Rousseeuw (1990) Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Son Inc.
19 - G. T. Heineman and W. T. Councill, editors. (2001) Component-based software engineering: putting the pieces together. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
20 - Clemens A. Szyperski. (1998) Component software - beyond object-oriented programming. Addison-Wesley-Longman.
21 -Raymond T. ,Jiawei Han. Eficient and Effective Clustering Methods for Spatial Data Mining
22- Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. International Conference on Knowledge Discovery and Data Mining

Datos de Contacto

NombrePol'la Matias Esteban
Institución:- GIISCO Research Group . Departamento de Ingeniería de Sistemas - Facultad de Informática
Universidad Nacional del Comahue - Neuquen, Argentina
Consejo Nacional de Investigaciones Cientícas y Técnicas – CONICET
Dirección:Buenos Aires 1400-(CP 8300) Neuquén- Argentina
email: matias.polla@fi.uncoma.edu.ar