FORMULACIÓN DE PROCESOS PARA UNA INGENIERÍA DE EXPLOTACIÓN DE INFORMACIÓN ESPACIAL

Tesista

Ing. Giovanni Daián ROTTOLI

Director

Rodolfo BERTONE (UNLP)

Asesor Científico

Hernán MERLINO (UNLA)

TESIS PRESENTADA
PARA OBTENER EL GRADO DE
DOCTOR EN CIENCIAS INFORMÁTICAS

FACULTAD DE INFORMÁTICA
UNIVERSIDAD NACIONAL DE LA PLATA

2022

RESUMEN

Los proyectos de explotación de información brindan un marco para la ejecución de actividades para la extracción de conocimiento a partir de conjuntos de datos para dar soporte a la toma de decisiones estratégicas. Esta extracción se realiza mediante la implementación de procesos de explotación de información que definen una secuencia lógica de actividades para la manipulación de datos y la búsqueda de patrones de conocimiento.

Entre los tipos de datos factibles a ser minados se encuentran los datos espacialmente referenciados que poseen un conjunto de características y propiedades que no son tenidas en cuenta por los procesos de explotación de información previamente definidos. Entre estas características se pueden mencionar la multiplicidad de tipos de representación, la variedad de relaciones espaciales implícitas que pueden ser extraídas a partir de las instancias de los mismos y que suelen ser valiosas en la toma de decisiones, y los fenómenos de autocorrelación y heterogeneidad espacial, que afectan la confianza en los resultados obtenidos.

Asimismo, la disponibilidad de herramientas software para el procesamiento de datos espaciales utilizadas en la implementación de los procesos es limitada, por lo que se busca posibilitar esta exploración mediante el uso de técnicas flexibles y presentes en multiplicidad de plataformas al día de la fecha.

En este contexto, el trabajo de tesis doctoral propone un conjunto de cuatro procesos de explotación de información espacialmente referenciada: un proceso para el descubrimiento de grupos espaciales, uno para el descubrimiento de anomalías, uno para el descubrimiento de asociaciones espaciales y uno para el descubrimiento de patrones de co-localización.

Estos procesos han sido diseñados considerando requerimientos no funcionales de utilidad, referida a la percepción de los usuarios sobre su uso para resolver problemas de negocios en contextos varios, interpretabilidad, relacionado con la facilidad de interpretación de los resultados, y adaptabilidad, considerando distintas formas de llevar

a cabo su implementación sobre distintas plataformas.

Finalmente, cada propuesta fue demostrada mediante el uso de distintos conjuntos de datos y sus requerimientos no funcionales fueron validados mediante el juicio de expertos, consolidando de esta forma un conjunto de estrategias para la resolución de problemas de negocios que involucran datos espacialmente referenciados considerando todas las particularidades que suceden a raíz de sus características contextuales.

ABSTRACT

Knowledge discovery processes provide a common framework for the execution of activities aimed to retrieve knowledge from datasets to support strategic decision-making. This knowledge retrieval is performed by implementing these knowledge discovery processes, which provide a logical sequence of activities for data manipulation in order to find patterns that contribute to finding the solution for problems on particular domains.

The possible data types to be mined include spatially-referenced data. This kind of data is linked to a set of properties and characteristics that are not taken into account by previously defined knowledge discovery processes. Some of these characteristics are the multiplicity of spatial data types; the variety of spatial relationships relevant for decision-making that are implicitly present among the instances of spatial objects; and, lastly, spatial autocorrelation and spatial heterogeneity, two properties that affect the trust in spatial data mining results.

Furthermore, the amount of software tools for spatial data processing and data mining is limited. Because of this, this work is also aimed at allowing pattern discovery by means of flexible tools that are present in multiple platforms.

In this context, this doctoral thesis proposes a set of four processes for spatial knowledge discovery: a process for discovery of spatial groups, a process for discovery of spatial outliers, a process for discovery of spatial associations, and a process for discovery of co-location patterns.

These processes have been designed taking into consideration non-functional requirements such as utility, which refers to the user's perceptions about the possibility of implementing them to solve business problems in different contexts; interpretability, related to the complexity of the outputs; and adaptability, considering different ways of implementing them on several platforms.

Lastly, each proposal was demonstrated using real-world datasets and the aforementioned requirements were validated considering expertsópinions. In consequence,

a set of new strategies to solve multiple business problems was consolidated, taking into account all the contextual characteristics related to spatial data.

AGRADECIMIENTOS

Un trabajo como el aquí presentado no se pudo haber sido llevado a cabo por una persona aislada. Son muchas las personas e instituciones que me han brindado su apoyo y ayuda de forma directa o indirecta y cabe dedicarles como mínimo unas líneas de agradecimiento.

En primer lugar, al profesor Dr. Ramón García-Martínez, quien fue mi mentor en los primeros años de este proyecto y me inició en el desarrollo de la investigación que ahora se materializa en esta tesis. Lamentablemente, agradecerle en persona ya no resulta posible. Su impronta no se pierde.

Luego, al profesor Dr. Hernán Merlino, quien me ha adoptado y guiado en su rol de asesor técnico y director de beca doctoral hasta la culminación exitosa de este trabajo.

Agradezco además a Facultad de Informática de la Universidad Nacional de La Plata, por nutrirme como *alma mater* y al Prof. Rodolfo Bertone de dicha casa de estudio por su rol de director de esta tesis doctoral.

A la Universidad Tecnológica Nacional, a su Facultad Regional Concepción del Uruguay, al departamento de Ingeniería en Sistemas de Información y a todos los que forman parte del mismo, por haber confiado en mis capacidades y haber permitido que crezca como docente e investigador.

Al Grupo de Investigación en Sistemas de Información (GISI) de la Universidad Nacional de Lanús y a todos sus investigadores por haberme dado un espacio entre ellos. En este contexto agradezco especialmente a Sebastián Martins y a Juan Manuel Rodríguez por su compañerismo durante estos años.

A los amables revisores que se han hecho un tiempo para leer estas páginas y regalarme sus oportunos comentarios. Sus apreciaciones han sido sumamente valoradas.

Un especial agradecimiento a Charly, al GIICIS y a la cátedra de Matemática Discreta, todos ellos de la Universidad Tecnológica Nacional, Facultad Regional Concepción del Uruguay, por haber sido compañeros de charlas y motor de las mejores ideas.

Por otro lado, en lo personal, agradezco a cada uno de mis amigos y a todos los que

han estado en mis momentos de debilidad. Estando cerca o lejos, todos han contribuido incansablemente con que sea un poco mejor cada día. Algunos ya no están presentes pero su huella ha quedado marcada profundamente en mi persona.

Agradezco en igual medida a mi familia, por haberme dado las alas necesarias para emprender el vuelo y las raíces para mantenerme firme.

Las palabras quedan cortas en cualquier sección de agradecimientos y este caso no es ajeno a ello. Resulta imposible abarcar a todos los que han formado parte de esta experiencia durante estos años pero confío en que todos ellos saben quienes son y en el valor de su ayuda. Por esto solo me queda reiterar: a todos y a cada uno de los que han contribuido directa e indirectamente con este proyecto, infinitamente gracias.

DEDICATORIA

"To absent friends, lost loves, old gods, and the season of mists;
and may each and every one of us always give the devil his due."

— Hob Gadling (The Sandman 2.22 - Neil Gaiman, 1991)

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ALGORITMOS

ÍNDICE DE DEFINICIONES

GLOSARIO

BI	Inteligencia de Negocio (*Business Intelligence*).
CLARANS	Algoritmo de agrupamiento *Clustering Large Applications based on Randomized Search)*
COD-CLARANS	Algoritmo de agrupamiento espacial considerando obstrucciones en el cálculo de la distancia.
COF	Algoritmo de detección de valores anormales que mejora el desempeño de LOF frente a escenarios particulares (*Connectivity-based Outlier Factor*).
CRISP-DM	Proceso Industrial Estándar para Minería de Datos (*Cross Industry Standard Process for Data Mining*).
DBSCAN	Algoritmo de agrupamiento basado en densidad (*Density-Based Spatial Clustering of Applications with Noise*).
DE-9IM	Modelo de 9 Intersecciones Dimensionalmente Extendido(*Dimensionally Extended nine-Intersection Model*).
EdI	Explotación de Información.
FSSG	Algoritmo de detección de subgrafos frecuentes.
GPS	Sistema de Posicionamiento Global (*Global Positioning System*).
GREW	Algoritmo de detección de subgrafos frecuentes.
HSIGRAM	Algoritmo de detección de subgrafos frecuentes.
IncGM+	Algoritmo de detección de subgrafos frecuentes.
IOWA	Operador de agregación ponderado con orden inducido *Induced Ordered Weighted Averaging Operator*
K-Means	Algoritmo de agrupamiento por particionamiento.
KDD	Proceso de Descubrimiento de Conocimiento en Bases de Datos (*Knowledge Discovery in Databases*).

LD-BSCA	Algoritmo de agrupamiento espacial basado en DBSCAN
LOF	Algoritmo de detección de valores anormales basado en densidades (*Local Outlier Factor*).
MLIOWA	Operador de agregación ponderado con orden inducido guiado por la mayoría *Majority guided induced Ordered Weighted Averaging Operator*
MoProPEI	Modelo de Proceso para Proyectos de Explotación de Información.
NSD(CLARANS)	Acercamiento espacialmente no dominante para el minado de datos espaciales (*Non-Spatial Dominant Approach using CLARANS*).
OWA	Operador de agregación ordenado y ponderado *Ordered Weighted Averaging Operator*
P3TQ	Modelo de Proceso para Proyectos de Explotación de Información.
SD(CLARANS)	Acercamiento espacialmente dominante para el minado de datos espaciales (*Spatial Dominant Approach using CLARANS*).
SEMMA	Modelo de Proceso para Proyectos de Explotación de Información.
SLOM	Algoritmo de detección de valores anormales basado en densidades considerando la vecindad espacial (*Spatial Local Outlier Measure*).
SOM	Mapas auto-organizados (*Self-Organizing Maps*), algoritmo de segmentación utilizando redes neuronales.
SUBDUE	Algoritmo de detección de subgrafos frecuentes.

CAPÍTULO 1

INTRODUCCIÓN

> "It is never 'only a dream', John Constantine. Here less than other places..."
>
> *Neil Gaiman, Dream a Little Dream of Me*

El primer capítulo de la tesis doctoral se estructura de la siguiente forma. En la sección 1.1 se presenta el contexto general en el cual se encuentra situada la tesis doctoral. Seguidamente, en la sección 1.2, se establece el objetivo de la misma. Luego se lista en la sección 1.3 la producción científica derivada de los resultados parciales de la investigación desarrollada. Por último, en la sección 1.4, se ofrece una visión general del contenido de este documento.

1.1. Contexto de la tesis

La Inteligencia de Negocio (BI, Business Intelligence), propone un enfoque interdisciplinario que contribuya a la generación de conocimiento para el soporte a la toma de decisiones estratégicas en las organizaciones, siendo una de estas disciplinas la Explotación de Información (EdI).

La EdI aporta desde el campo de la informática herramientas de análisis y síntesis de información para la extracción de conocimiento no trivial e implícito en los datos presentes en los repositorios de las organizaciones.

Entre estas herramientas se destacan los procesos de explotación de información,

que estructuran de una forma lógica las actividades a ser llevadas a cabo sobre los conjuntos de datos, de manera tal que las regularidades a descubrir sean valiosas para la resolución de los problemas de los tomadores de decisiones (*stakeholders*).

Por otro lado, los datos espacialmente referenciados son una clase particular de datos dependientes del contexto, los cuales permiten modelar eventos, objetos o situaciones localizados en un espacio determinado.

Utilizando esta clase de datos como entrada es posible extraer patrones de comportamiento que no sólo involucren a los atributos que lo componen, sino también a las interacciones espaciales que ocurren entre cada uno de ellos, como pueden ser relaciones de vecindad, de posicionamiento, entre otras. Esto aumenta la complejidad de obtención de los mismos en comparación con la minería de datos y explotación de información tradicional.

No obstante, si bien existen algoritmos que permiten extraer estas relaciones complejas, se puede observar en la literatura que se suelen abordar problemas de negocios relacionados con este tipo de datos ignorando múltiples aspectos conceptuales que agregan valor a los patrones extraídos. La falta de conocimiento específico, la falta de herramientas software que implementen los métodos necesarios o la falta de una estructura base que guíe el proceso según los distintos intereses de los tomadores de decisiones, son posiblemente los principales motivadores que inducen este problema.

De esta forma, por lo general, se aplican algoritmos de minería de datos tradicionales siguiendo procesos estructurados no pensados para manejar las peculiaridades asociadas a este tipo de información contextual.

Por consiguiente, se plantean las siguientes hipótesis de investigación:

H1 Las propiedades particulares de la información espacialmente referenciada aumentan la complejidad en la obtención de patrones de conocimiento de interés para la inteligencia de negocio.

H2 Una gran cantidad de abordajes prácticos utilizados para la extracción de conocimiento a partir de datos espaciales ignoran las propiedades de los mismos, lo cual impacta en el correcto uso de los patrones de conocimiento hallados.

H3 Las herramientas tecnológicas actuales no resultan de fácil acceso para los analistas, por lo que se dificulta el prototipado de las soluciones.

1.2. Objetivo de la tesis

En el contexto mencionado con anterioridad, se plantea como objetivo general del trabajo de investigación desarrollar y validar Procesos de Explotación de Información Espacial que sistematicen la obtención de conocimiento sin ignorar las propiedades particulares de este tipo de datos.

En virtud de lo expuesto, resulta imperioso identificar las necesidades de los posibles *stakeholders* en lo que se refiere a descubrimiento de conocimiento en bases de datos espaciales; y de esta manera, diseñar procesos consistentes en actividades de manipulación y analítica de datos para dar respuesta a las mismas.

De esta manera y ante situaciones particulares, un proceso puede ser instanciado de acuerdo a las características particulares del dominio del problema y de los datos de entrada, haciendo uso de distintas herramientas informáticas para extraer los patrones de conocimiento adecuados para tal caso.

El diseño de estos procesos ha de tener en cuenta cuáles son las características que se asocian con la estructura de los datos y bases de datos espaciales. Además, se deberá considerar la posibilidad de representar distintas relaciones espaciales entre los datos, así como también la posible ocurrencia de fenómenos asociados con la localidad de los objetos, tales como la heterogeneidad espacial y la autocorrelación espacial.

Particularmente, se plantean los siguientes objetivos específicos:

- Identificar los tipos de patrones de conocimiento que se pueden obtener a partir de los datos espaciales que resulten de interés para la inteligencia de negocios.
- Identificar las propiedades de los datos espaciales que aportan valor para la toma de decisiones estratégicas en relación a cada patrón de conocimiento particular.
- Diseñar procesos de explotación de información para la obtención de patrones de conocimiento a partir de datos espaciales haciendo uso de técnicas y soluciones implementadas en las herramientas tecnológicas actuales, de forma tal que sean lo suficientemente flexibles para ser adaptados a distintos escenarios particulares.

1.3. Producción científica derivada

A continuación se presentan los resultados parciales comunicados en revistas internacionales (sección 1.3.1), simposios doctorales (sección 1.3.2), conferencias de alcance internacional (sección 1.3.3), conferencias de alcance nacional (sección 1.3.4) y capítulos de libros (sección 1.3.5). Todas las publicaciones han sido sometidas a un proceso de evaluación por pares.

1.3.1. Revistas internacionales con referato

- Rottoli, G.D. y Merlino H. (2020). Spatial Association Discovery Process using Frequent Subgraph Mining. *TELKOMNIKA (Telecommunication, Computing, Electronics and Control)* 18.4. `http://dx.doi.org/10.12928/telkomnika.v18i4.13858`

1.3.2. Simposios doctorales

- Rottoli, G.D. (2018). Formulación de Procesos para una Ingeniería de Explotación de Información Espacial.*XXI Ibero-american Conference on Software Engineering* (CIBSE 2018). Doctoral Symposium. Bogotá, Colombia.

1.3.3. Conferencias internacionales con referato

- Rottoli, G. D., Merlino, H., & García-Martinez, R. (2017). Co-location Rules Discovery Process Focused on Reference Spatial Features Using Decision Tree Learning. *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (pp. 221-226). Springer, Cham. `https://doi.org/10.1007/978-3-319-60042-0_25`

- Rottoli, G. D., Merlino, H., & García-Martinez, R. (2017). Knowledge discovery process for description of spatially referenced clusters. *International Conference on Software Engineering & Knowledge Engineering*. Ed. USA KSI Research Inc. and Knowledge Systems Institute (Vol. 410415). `https://doi.org/10.18293/SEKE2017-013`

- Rottoli, G. D., Merlino, H., & García-Martínez, R. (2018). Knowledge Discovery Process for Detection of Spatial Outliers. *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (pp. 57-68). Springer, Cham. https://doi.org/10.1007/978-3-319-92058-0_6

1.3.4. Conferencias nacionales con referato

- Rottoli, G. D., Merlino, H., & García Martínez, R. (2016). Proceso de descubrimiento de reglas de caracterización de grupos espacialmente referenciados. *XXII Congreso Argentino de Ciencias de la Computación* (CACIC 2016).

- Rottoli, G. D., Merlino, H., & García Martínez, R. (2016). Proceso de descubrimiento de patrones de co-localización alrededor de tipos de eventos de referencia. *XXII Congreso Argentino de Ciencias de la Computación* (CACIC 2016).

- Rottoli, G. D., Merlino, H., & García Martínez, R. (2018). Proceso para el Descubrimiento de Asociaciones Espaciales Mediante Subgrafos Frecuentes. *XXIV Congreso Argentino de Ciencias de la Computación* (CACIC 2018).

- Rottoli, G. D., Merlino, H. (2019). Extensión de Proceso de Explotación de Información para el Descubrimiento de Asociaciones Espaciales Difusas. *XXV Congreso Argentino de Ciencias de la Computación* (CACIC 2019).

- Rottoli, G. D., Merlino, H. (2020). Proceso para la Obtención de Patrones de Co-localización con Relaciones Difusas. *Simposio Argentino de Inteligencia Artificial. Jornadas Argentinas de Informática* (ASAI-JAIIO, 2020) (Short paper).

1.3.5. Capítulos de libros

- Rottoli, G. D., Merlino, H., & García-Martinez, R. (2017). Discovery Process of Co-Localization Patterns around Reference Event Types. *Computer Science & Technology Series: XXII Argentine Congress of Computer Science Selected Papers* (pp. 183-190). https://doi.org/10.35537/10915/61164

1.4. Visión general de la tesis

El presente documento se estructura en siete capítulos: “Introducción”, “Estado de la Cuestión”, “Descripción del Problema”, “Solución”, “Validación”, “Conclusiones” y “Referencias”.

En el primer capítulo, Introducción, se plantea el contexto de la tesis doctoral y sus objetivos así como también las publicaciones en revistas científicas, simposios doctorales y conferencias de alcance nacional e internacional que validan los resultados parciales del trabajo realizado.

El segundo capítulo, Estado de la Cuestión, describe el marco teórico sobre el que se erige el presente trabajo. Se presentan los conceptos de Explotación de Información y Minería de Datos detallando metodologías para el desarrollo de proyectos en esta disciplina. Además, se profundiza en el concepto de proceso de explotación de información, clave para la formulación de los objetivos de la investigación. Por último, se enumeran y se detallan las características de los datos espacialmente referenciados y de la minería de este tipo particular de información poniéndola en contraste con las características de las bases de datos relacionales o transaccionales.

Luego, el capítulo 3 titulado “Descripción del problema”, detalla la problemática derivada del estado del arte que ha sido tratada en el contexto de la investigación doctoral, presentando formalmente los objetivos y preguntas de investigación que guían el trabajo.

Posteriormente, el capítulo 4, Solución, presenta inicialmente el conjunto de herramientas y teorías utilizadas para la construcción de las soluciones propuestas para el problema planteado en el capítulo precedente. Luego se describen cada una de estas soluciones y se presenta en las últimas páginas de esta sección una adaptación difusa de los procesos, particularmente aquellos en los que intervienen de forma más explícita diferentes clases de relaciones espaciales.

En el capítulo 5, Validación, se utilizan los procesos planteados para resolver problemas en contextos específicos en los que los datos espaciales son clave para su resolución. Aquí se instancian cada una de las actividades propuestas en el capítulo anterior haciendo uso de tecnologías de fácil acceso. Estas implementaciones son usadas como entrada de un proceso de evaluación por juicio de expertos. Sus opiniones son analizadas y agregadas utilizando teoría de conjuntos difusos.

En el capítulo 6, se presentan los aportes de esta tesis doctoral y se destacan las futuras líneas de investigación relacionadas con el trabajo realizado y las cuestiones pendientes que exceden el alcance de la tesis.

Finalmente, el capítulo “Referencias” presenta las publicaciones consultadas que soportan el desarrollo de la tesis.

CAPÍTULO 2

ESTADO DE LA CUESTIÓN

> "Do you know how long it's been since I mislaid a book? Well, let's just say the continents weren't in their current shapes, not that that means anything to you."
>
> *Neil Gaiman, The Hunt*

El presente capítulo introduce la temática abordada en la tesis doctoral y presenta el marco teórico en el que se sustenta la misma. Primeramente, en la sección 2.1 se definen los conceptos relacionados con la explotación de información y la minería de datos. En la sección 2.2 se detallan metodologías para el desarrollo de proyectos de explotación de información. Luego, en la sección 2.3 se describen brevemente los procesos de explotación de información. Posteriormente, se hace hincapié en los datos espacialmente referenciados y en sus características distintivas en la sección 2.4, mientras que en la sección 2.5 se profundiza en las relaciones posibles entre estos objetos. En las secciones 2.6 y 2.7 se describen dos fenómenos que atañen a la contextualización de los datos en el espacio: la autocorrelación y la heterogeneidad espacial. Por último, en la sección 2.8 se mencionan actividades de minería de datos y sus particularidades al trabajar con esta clase de datos de entrada.

2.1. Explotación de información

Se denomina Inteligencia de Negocio (BI, del inglés *Business Intelligence*) al conjunto de herramientas y métodos que apoyan la generación de conocimiento para la toma de decisiones estratégicas en una organización (Rud, 2009).

La Explotación de Información (EdI) es una disciplina de la informática que brinda a la inteligencia de negocio herramientas de análisis necesarias para la extracción de conocimiento no trivial que se encuentra de forma implícita en repositorios de información (García-Martínez et al., 2011).

Este conocimiento es obtenido a partir de la búsqueda de regularidades interesantes entre los objetos contenidos en los repositorios mencionados mediante el uso de algoritmos de minería de datos, los cuales provienen generalmente de campos de estudio como la inteligencia artificial y el aprendizaje automático (García-Martínez et al., 2013).

La aplicación de un enfoque sistemático al desarrollo de proyectos de explotación de información define lo que se conoce como "ingeniería en explotación de información", la cual establece un conjunto de procedimientos que guían el desarrollo de estas actividades con el fin de garantizar que la producción de conocimiento satisfaga las necesidades reales de los interesados.

Los proyectos de explotación de información permiten asegurar la calidad final de los resultados obtenidos al establecer controles sobre cada una de las etapas que intervienen en el proceso productivo, coexistiendo las actividades técnicas con las de gestión del proyecto y las propias de la organización que lo ejecuta (Pollo Cattaneo et al., 2010).

Es menester, además, fundar este enfoque sistemático con una estructura base que establezca las principales actividades tanto de características técnicas para el desarrollo del proyecto, como de gestión mismo. Esta estructura se conoce como "modelo de proceso para proyectos de explotación de información" (Martins et al., 2016), cuestión que es profundizada en la sección siguiente.

2.2. Modelos de procesos para proyectos de explotación de información

En la historia de la explotación de información han surgido distintos acercamientos metodológicos para estructurar los proyectos de explotación de información de una forma sistematizada. Entre estas propuestas destacan el proceso KDD (*Knowledge Discovery in Databases*) (Fayyad et al., 1996), CRISP-DM (*Cross Industry Standard Process for Data Mining*)(Chapman et al., 1999), SEMMA (Azevedo and Santos, 2008), P3TQ (Pyle, 2003) y MoProPEI (Modelo de Procesos para Proyectos de Explotación de Información) (Martins et al., 2016).

Cada una de estas herramientas permiten hacer énfasis en distintos aspectos de los proyectos de explotación de información debido a sus características particulares, incluyendo fortalezas y debilidades propias.

En las secciones siguientes del presente documento se pondrá especial atención en KDD, por haber sido la primera propuesta que diferencia los conceptos "minería de datos" y "explotación de información", en CRISP-DM, por ser el estándar *de facto* para llevar a cabo proyectos de explotación de información desde su lanzamiento en el año 1999 (Schröer et al., 2021), y en MoProPEI, por ser una propuesta novedosa que no sólo se enfoca en el descubrimiento de patrones de utilidad para la toma de decisiones sino que además incorpora actividades de gestión para asegurar el éxito del proyecto.

2.2.1. Proceso KDD

El proceso de descubrimiento de conocimiento en bases de datos (KDD, del inglés Knowledge Discovery in Databases) busca sistematizar el uso de algoritmos de minería de datos y la selección de hipótesis. Este proceso, propuesto por Fayyad et al. (1996), diferencia los conceptos "minería de datos" y "explotación de información", entendiendo el primero como la actividad de aplicar algoritmos sobre datos para la búsqueda de regularidades en los mismos, y el segundo como el proceso necesario para la obtención de conocimiento a partir de dichos repositorios. Consecuentemente, la minería de datos resulta un elemento constitutivo del proceso de explotación de información.

El proceso KDD se construye sobre un entendimiento del dominio de aplicación,

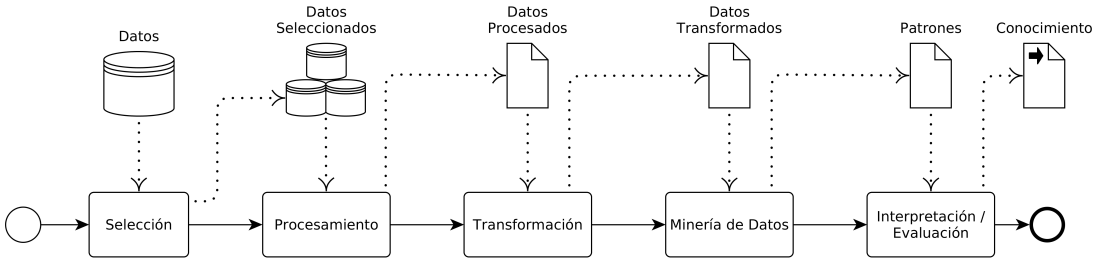


Figura 2.1: Proceso de Descubrimiento de Conocimiento en Base de Datos (KDD).

recolección de conocimiento previo y la identificación de los objetivos desde la visión de los *stakeholders*.

Como detalla la figura 2.1, en el primer paso del proceso se determina cuál es el conjunto de datos a ser utilizado mediante la selección de ejemplos y atributos. Luego, estos datos son limpiados y preprocesados para aumentar la calidad de los mismos mediante el tratamiento de ruido, anomalías, datos perdidos, errores, entre otros. Posteriormente, si la situación lo amerita, se construyen nuevos atributos a partir de los anteriores mediante técnicas de reducción de la dimensionalidad, discretización, normalización, aplicación de funciones de agregación, etc. En el anteúltimo paso se extraen regularidades implícitas en el conjunto de datos mediante el uso de algoritmos de minería de datos. Para finalizar, se interpretan los resultados para generar conocimiento útil para la toma de decisiones.

2.2.2. CRISP-DM

La metodología CRISP-DM, del inglés *Cross Insdustry Standard Process for Data Mining*, es un modelo de proceso o metodología estándar para el desarrollo de proyectos de explotación de información, independiente de herramientas, técnicas o incluso de su dominio de aplicación (Chapman et al., 1999).

CRISP-DM se organiza en cuatro niveles jerárquicos que van desde un nivel abstracto general hasta niveles concretos de implementación para un problema determinado. Estos niveles, tal como se observa en la figura 2.2, son 1. Fases, 2. Actividades Genéricas, 3. Actividades Específicas, y 4. Instancias de Proceso.

El nivel de fase es el más general y abstracto de la metodología. Este define seis componentes denominadas “fases” que se organizan en un ciclo de vida iterativo con retroalimentación entre algunas de ellas. La figura 2.3 muestra las interacciones de

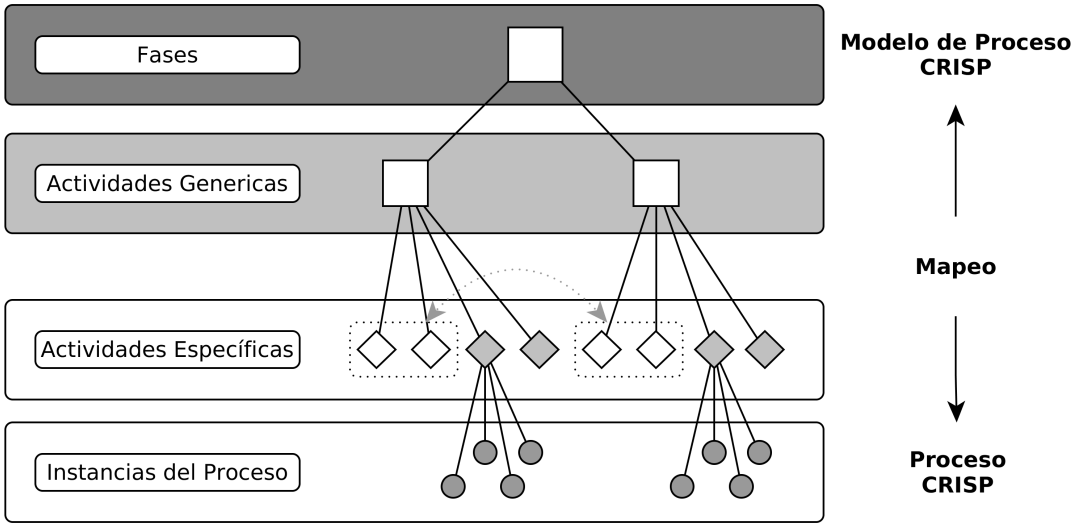


Figura 2.2: Jerarquía de niveles de CRISP-DM.

cada fase dentro de este ciclo de vida. A continuación, en las siguientes secciones se profundiza en las mismas.

2.2.2.1. Comprensión del negocio

A diferencia del proceso KDD, CRISP-DM aporta una visión del contexto en la que se desarrollan las actividades de minería de datos. Se mueve el foco desde las actividades que son desarrolladas para la extracción de patrones de conocimiento a la satisfacción de objetivos de negocio bien establecidos al comienzo del proyecto.

La primera fase de la metodología, comprensión del negocio (*business understanding*) tiene como propósito el establecimiento de los objetivos del negocio y los objetivos de explotación de información, así como también el desarrollo de la planificación para la ejecución del proyecto.

Esta fase se desglosa en cuatro actividades generales (tabla 2.1): 1. Determinar los objetivos del negocio, para establecer las necesidades del cliente y los criterios de éxito del proyecto; 2. evaluar la situación actual, para analizar los recursos disponibles, los requisitos y los riesgos asociados a la ejecución del proyecto; 3. determinar los objetivos de minería de datos, para definir el conocimiento necesario de ser extraído para satisfacer los objetivos de negocio planteados; 4. planificar el proyecto, para definir los pasos futuros y las herramientas tecnológicas a ser utilizadas.

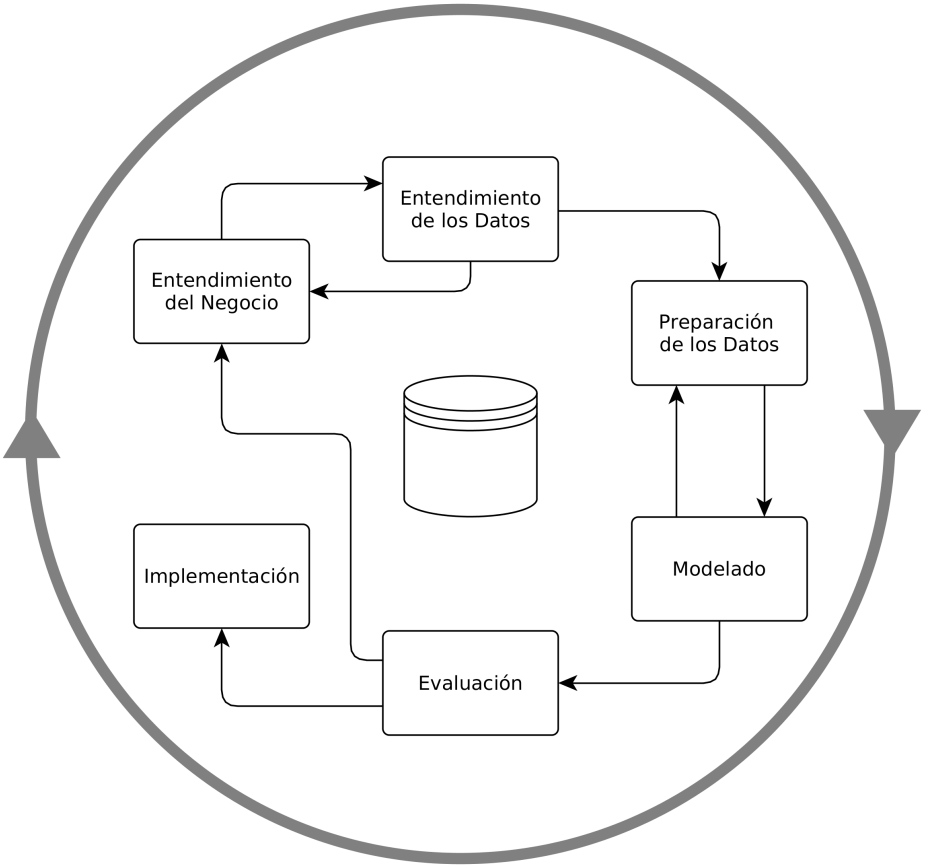


Figura 2.3: Ciclo de vida de CRISP-DM.

Tabla 2.1: Actividades generales de la fase de comprensión del negocio de CRISP-DM.

Actividad general	**Salidas**
Determinar objetivos del negocio	· Antecedentes · Objetivos de negocio · Criterios de éxito
Evaluar de la situación actual	· Inventario de recursos · Requisitos, supuestos y restricciones · Riesgos y contingencias · Terminología · Costos y beneficios
Establecer objetivos de requisito	· Objetivos de minería de datos · Criterios de éxito de minería de datos
Producir el plan del proyecto	· Planificación del proyecto · Evaluación de herramientas y técnicas

Tabla 2.2: Actividades generales de la fase de comprensión de los datos de CRISP-DM.

Actividad general	Salidas
Recolectar los datos	· Reporte de la colección de datos iniciales
Describir los datos	· Reporte de la descripción de los datos
Explorar los datos	· Reporte de exploración de los datos
Verificar calidad de los datos	· Reporte de calidad de los datos

2.2.2.2. Comprensión de los datos

La comprensión de los datos (*data understanding*) es la segunda fase de CRISP-DM. Esta etapa tiene como objetivo conocer inicialmente la materia prima del proyecto de explotación de información: los datos. Una vez hecho esto, podremos definir la estrategia para poder solucionar los problemas de negocio planteados en la fase anterior.

Las actividades generales de esta fase son 1. recolección inicial de datos, donde se adquiere acceso a los repositorios de la organización interesada; 2. descripción de los datos, donde se adquiere conocimiento sobre la constitución de los repositorios adquiridos, sus formatos, atributos, entre otros; 3. exploración de los datos, para detectar particularidades de comportamiento iniciales de los atributos, con el fin de hallar características interesantes que puedan contribuir con el logro de los objetivos; 4. verificación de la calidad de datos, que determinará si se cumplen condiciones mínimas en cuanto a completitud, errores, u omisiones de valores de atributos (tabla 2.2).

En caso de detectarse algún problema con los datos, como la imposibilidad de uso por no cumplir con la calidad necesaria, es posible regresar a la fase anterior para re-evaluar los alcances del proyecto.

2.2.2.3. Preparación de los datos

Luego de la obtención de la descripción de los datos adquiridos que nutren el proyecto de explotación de información, se prosigue con la preparación de los datos (*data preparation*) para adecuarlos a las técnicas y herramientas de modelado que serán ejecutadas en la próxima fase.

Esta adecuación se realiza mediante las siguientes actividades generales (tabla 2.3): 1. especificar el conjunto de datos, cuyo propósito es describir el conjunto de datos que será utilizado posteriormente en la fase de modelado; 2. selección, en la que se

Tabla 2.3: Actividades generales de la fase de preparación de los datos de CRISP-DM.

Actividad general	**Salidas**
Seleccionar datos	· Criterios de inclusión/exclusión
Limpiar datos	· Reporte de limpieza de datos
Construir datos	· Atributos derivados
	· Registros generados
Integrar datos	· Datos combinados
Formatear datos	· Datos formateados

opta por ciertas tuplas y atributos que son relevantes según los objetivos planteados y que cumplen con las condiciones necesarias de calidad; 3. limpieza, para aumentar la calidad de los datos mediante la eliminación de tuplas u valores erróneos de la base de datos; 4. construcción, que incluye la obtención de atributos derivados que aporten directamente al conocimiento que permite el logro de los objetivos; 5. integración, para combinar atributos de distintos repositorios, internos o externos; 6. formateo, mediante modificaciones realizadas sobre los datos, sin cambiar su semántica, para adecuarlos a las técnicas y herramientas de modelado disponibles.

2.2.2.4. Modelado

En la fase de modelado (*modeling*) se aplican los distintos algoritmos y técnicas destinadas a la búsqueda de regularidades en los datos: los algoritmos de minería de datos.

Para esto la fase se desglosa en cuatro actividades generales (tabla 2.4): 1. selección de técnica de modelado, la cual está dada por los objetivos de minería de datos planteados y las características de los datos preparados en la fase anterior; 2. generación de la prueba de diseño, que planificará la ejecución de los entrenamientos de los modelos seleccionados y evaluación de los resultados; 3. construcción del modelo, en la que se generarán los modelos predictivos o descriptivos seleccionados según el plan especificado; 4. evaluación del modelo, en la que se ejecutarán las pruebas de eficiencia y precisión de los modelos generados en la actividad anterior.

En caso de ser necesaria una modificación del conjunto de datos por cuestiones técnicas, es factible regresar a la fase anterior: preparación de los datos.

Tabla 2.4: Actividades generales de la fase de modelado de CRISP-DM.

Actividad general	Salidas
Elegir técnicas de modelado	· Técnicas de modelado · Supuestos de modelado
Generar diseño de pruebas	· Diseño de las pruebas
Construir modelos	· Configuración de parámetros · Modelos · Descripción de los modelos
Evaluar modelos	· Evaluación de los modelos · Configuración revisada de parámetros

Tabla 2.5: Actividades generales de la fase de evaluación de CRISP-DM.

Actividad general	Salidas
Evaluar resultados	· Evaluación de los resultados con respecto a los criterios de éxito de minería de datos · Modelos aprobados
Revisar el proceso	· Revisión del proceso
Determinar próximos pasos	· Lista de posibles acciones · Decisión

2.2.2.5. Evaluación

En la fase de evaluación (*evaluation*), se determina si los patrones de conocimiento obtenidos en la fase anterior satisfacen los objetivos planteados al comienzo del proyecto y los criterios de éxito establecidos. En caso de no satisfacer las condiciones iniciales, será necesario re-evaluar el proyecto desde el inicio para hallar errores conceptuales o técnicos.

Esta fase comprende las actividades generales de 1. evaluación de resultados, en la que se determina si el conocimiento satisface los criterios de éxito que constan en la especificación del proyecto; 2. revisión del proyecto, en la que se determina si existen actividades omitidas o cuya ejecución no fue 100 % exitosa, con el fin de abordarlas nuevamente en la próxima iteración del ciclo de vida; 3. determinación de los próximos pasos, donde se decide si finalizar el proyecto, tras lo cual se procedería a la fase siguiente, o reiterar desde el comienzo (tabla 2.5).

Tabla 2.6: Actividades generales de la fase de implementación de CRISP-DM.

Actividad general	Salidas
Planificar implementación	· Plan de implementación
Planificar vigilancia y mantenimiento	· Plan de mantenimiento
Producir reporte final	· Reporte final
	· Presentación final
Revisar proyecto	· Documentación de experiencia

2.2.2.6. Implementación

La última fase, implementación (*deployment*), tiene como objetivo la generación del plan de desarrollo, incluyendo la documentación y presentación de los resultados del cliente.

Las cuatro tareas generales de esta fase son (tabla 2.6): 1. realizar el plan de implementación, que se alimenta de los resultados de la evaluación para determinar una estrategia para la presentación o uso de los mismos; 2. realizar el plan de vigilancia y mantenimiento, para proyectar las actividades que se asegurarán de que los resultados obtenidos son relevantes en el tiempo y se encuentran actualizados; 3. realizar el informe final; 4. revisar el proyecto, identificando posibles mejoras y las experiencias que pueden ser útiles en futuros proyectos.

2.2.2.7. KDD y CRISP-DM

Tal como se adelantó en la sección previa, el proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD) hace foco en la manipulación de datos: su selección, preprocesamiento y aplicación de algoritmos de minería de datos para la obtención de patrones de conocimiento.

El proceso CRISP-DM, por otro lado, agrega al primero actividades dedicadas a la comprensión del dominio del problema, el establecimiento y formalización de los requisitos y objetivos, y la elaboración de planificaciones para conducir el desarrollo del proyecto. Estas actividades no son consideradas formalmente dentro del proceso KDD, mas son tenidas en cuenta como conocimiento previo, tal como se detalló anteriormente.

Asimismo, CRISP-DM agrega una etapa de evaluación de los resultados que no posee KDD, en la cual se analiza si el conocimiento obtenido contribuyen al logro de

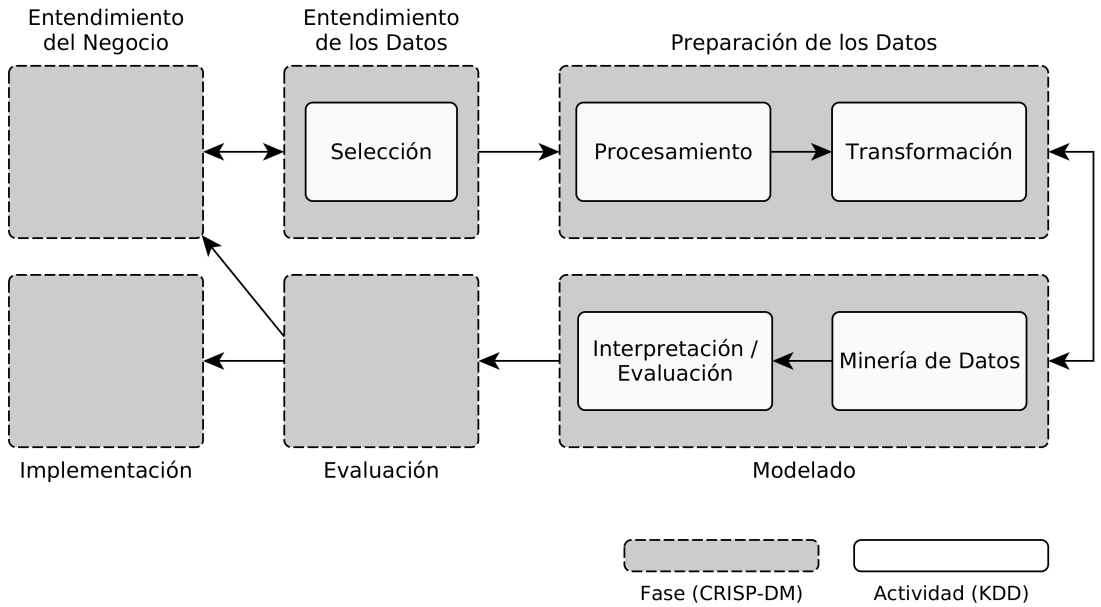


Figura 2.4: Comparación de actividades de KDD y CRISP-DM.

los objetivos planteados.

Como se muestra en la figura 2.4, se puede interpretar KDD como un subproceso de CRISP-DM, donde la fase de Selección de KDD se puede corresponder con ciertas actividades de la fase de Entendimiento de los Datos de CRISP-DM, las fases de Procesamiento y de Transformación del primero con la fase de Preparación de los Datos del segundo, y las etapas de Minería de Datos y Evaluación/Interpretación pueden encontrarse dentro de la etapa de Modelado de la metodología estándar (Azevedo and Santos, 2008).

2.2.3. MoProPEI

El Modelo de Proceso para Proyectos de Explotación de Información (MoProPEI), propuesto por Martins et al. (2016), brinda un marco de gestión transversal para el desarrollo de proyectos EdI, enfocándose en remover huecos existentes en metodologías como CRISP-DM, aumentando así la madurez y la calidad de los proyectos.

Esta propuesta se compone de dos subprocesos: un subproceso de gestión y un subproceso de desarrollo. En primer lugar, el subproceso de desarrollo se encarga de la generación del conocimiento de forma similar al proceso CRISP-DM, mientras que el subproceso de gestión lo atraviesa para gestionar y controlar todo el proyecto hacien-

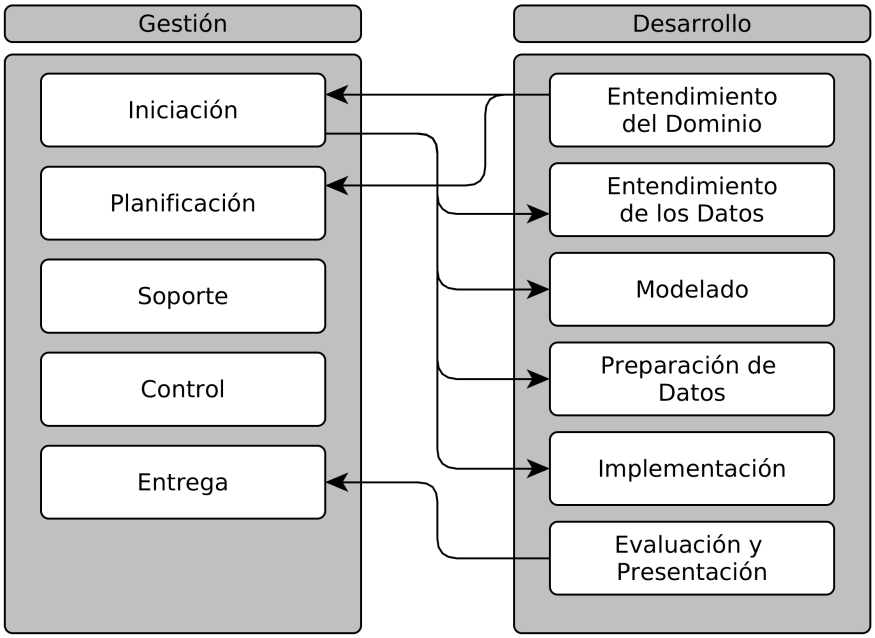


Figura 2.5: Estructura de fases de MoProPei.

do uso de herramientas de estimación de esfuerzos, de validación (Pytel et al., 2013, 2015), entre otras.

Como se puede observar en la figura 2.5, el subproceso de gestión está conformado por 5 fases que son detalladas en las siguientes subsecciones. Estas fases se relacionan a su vez con las fases del subproceso de desarrollo. Para mayores detalles al respecto consultar (Martins, 2020)

2.2.3.1. Iniciación

La fase de Iniciación es la primer a fase del subproceso de gestión de MoProPEI; contiene cuatro actividades: 1. comunicar el protocolo, que busca determinar los canales formales de interacción de acuerdo a las posibilidades del equipo de trabajo y los expertos del dominio del problema; 2. explorar conceptos iniciales, que permite que el equipo de trabajo tenga una primera aproximación a las características del cliente y a sus necesidades; 3. evaluar situación, que analiza las características del proyecto, los requerimientos del cliente, los posibles riesgos y sus contingencias, y demás particularidades para evaluar la factibilidad del mismo; 4. definir ciclo de vida, que determina la estructura y las iteraciones del proyecto.

2.2.3.2. Planificación

La fase que continúa es la fase de Planificación, que se encuentra compuesta por tres actividades principales: 1. planificar de actividades, cuyo propósito es determinar y organizar el tiempo necesario para finalizar cada actividad del proyecto, produciendo como salida un calendario y una lista de métricas; 2. planificar recursos, que busca determinar qué recursos y en qué momento serán requeridos; 3. estimaciones y responsabilidades, que determina la duración y costo del proyecto, así como también el alcance del mismo y las obligaciones de las partes.

2.2.3.3. Soporte

La tercera fase, Soporte, está compuesta por tres actividades: 1. gestión del ciclo de vida, tiene como objetivo formalizar el alcance de cada iteración del proyecto y establecer los logros alcanzados, reajustando el alcance en cada iteración si fuere necesario; 2. gestión de la implementación, cubre aquellas actividades relacionadas con el desarrollo del producto de explotación de información, tales como definir los objetivos de cada paso del proyecto, integrando los productos internos, comunicando el progreso del mismo, etc.; 3. gestión de la configuración, cuyo objetivo es documentar la información relevante producida durante el proyecto.

2.2.3.4. Control y calidad

La fase siguiente, denominada “Control y Calidad”, involucra cuatro actividades orientadas a la mejora de la madurez del proyecto de explotación de información y de sus productos, manteniendo un seguimiento de todos sus partes constituyentes: 1. control de los recursos, busca verificar el cumplimiento del plan de proyecto sin considerar la incorporación de recursos; 2. mediciones, busca cuantificar los elementos de interés para realizar un análisis cuantitativo del proyecto, usando las métricas definidas previamente; 3. control de actividades, evalúa características del proyecto como el progreso de las actividades, la duración y costo de las mismas, los riesgos, la calidad, entre otras, para identificar posibles desviaciones que obstaculicen el cumplimiento de los objetivos propuestos; 4. control del cambio, cubre todas las actividades cuyo objetivo sea evaluar, implementar y monitorizar los cambios requeridos por el proyecto.

2.2.3.5. Entrega

Por último, la fase de Cierre Formal o Entrega se encuentra conformada por dos actividades: 1. cierre formal externo, que se encarga de asegurar el logro de los objetivos requeridos por el cliente; 2. cierre formal externo, cuyo objetivo es obtener conocimiento a partir de las características analizadas durante la ejecución del proyecto, útil para la ejecución de futuros proyectos.

2.3. Procesos de explotación de información

Un proceso de explotación de información se define como un conjunto de actividades lógicamente relacionadas que se ejecutan para obtener a partir de un conjunto de información con cierto grado de valor para la organización en un contexto dado, otro conjunto de información con un grado de valor mayor que el inicial (Ferreira et al., 2005).

Britos (2008) propone un conjunto de procesos de explotación de información basados en sistemas inteligentes con el propósito de ser aplicados para dar respuesta a problemáticas específicas de la inteligencia del negocio (García-Martínez et al., 2013). Estos procesos son:

- **Proceso de descubrimiento de reglas de comportamiento:** (Fig. 2.6a) aplica cuando se requiere identificar cuáles son las condiciones para obtener determinados resultados en el dominio del problema. Son ejemplos de problemas que requieren este proceso la identificación de las características del local más visitado por los clientes, la identificación de factores que inciden en el alza de las ventas de un producto dado, el establecimiento de las características o rasgos de los clientes con alto grado de fidelidad a la marca, entre otros.

- **Proceso de descubrimiento de grupos:** (Fig. 2.6b) aplica cuando se requiere identificar una partición en la masa de información disponible sobre el dominio de un problema. Son ejemplos de problemas que requieren este proceso la identificación de segmentos de clientes para bancos y financieras, la identificación de tipos de llamadas de los clientes para empresas de telecomunicación, la identificación de grupos sociales con las mismas características, etc.

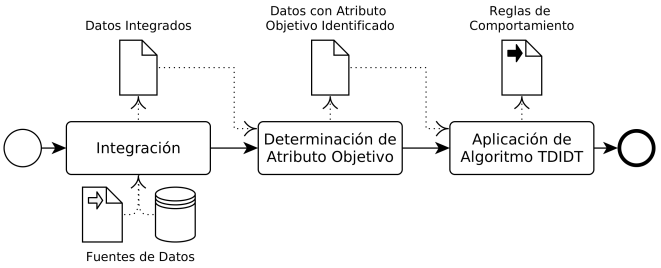
- **Proceso de ponderación de interdependencia de atributos:** (Fig. 2.6c) El proceso de ponderación de interdependencia de atributos aplica cuando se requiere identificar cuáles son los factores con mayor incidencia (o frecuencia de ocurrencia) sobre un determinado resultado de un problema. Entre otros, son ejemplos de problemas de aplicabilidad de este proceso la determinación de factores que poseen incidencia sobre las ventas y la determinación de los rasgos distintivos de clientes con alto grado de fidelidad a la marca.

- **Proceso de descubrimiento de reglas de pertenencia a grupos:** (Fig. 2.6d) se utiliza cuando se busca identificar cuáles son las condiciones de pertenencia a cada una de las clases en una partición desconocida "a priori", pero que se encuentra presente en la masa de información disponible sobre el dominio de problema. Son ejemplos de problemas que requieren este proceso: el establecimiento de la tipología de perfiles de clientes y la caracterización de cada tipología, la distribución y estructura de datos de un website, entre otros.

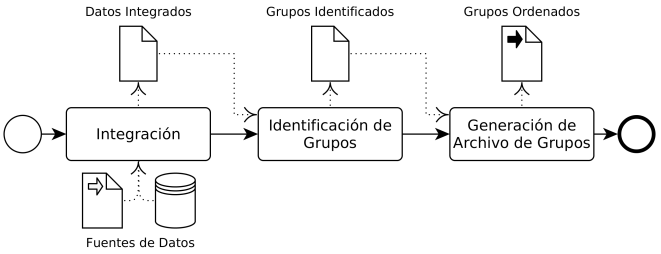
- **Proceso de ponderación de reglas de comportamiento o de reglas de pertenencia a grupos:** (Fig. 2.6e) es de utilidad cuando se requiere identificar las condiciones con mayor incidencia sobre la obtención de un determinado resultado en el dominio del problema, sean éstas las que en mayor medida inciden sobre un comportamiento o las que mejor definen la pertenencia a un grupo. Son ejemplos de problemas que requieren este proceso: la identificación del factor dominante que incide en el alza las ventas de un producto dado, el rasgo con mayor presencia en los clientes con alto grado de fidelidad a la marca, la frecuencia de ocurrencia de cada perfil de clientes, la identificación del tipo de llamada mas frecuente en una región, etc.

2.4. Datos espaciales

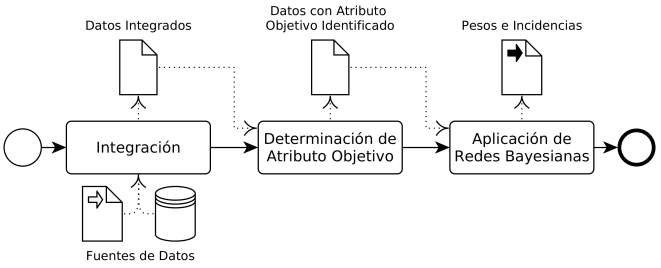
En los últimos años se ha evidenciado un crecimiento explosivo de la cantidad de datos asociados al espacio, principalmente geográfico. Este crecimiento se debe al surgimiento de tecnologías como el sistema de posicionamiento global (Global Positioning System, GPS), la telefonía celular y las imágenes censadas remotamente mediante satélites y vehículos no tripulados (drones), entre otros (Golmohammadi et al., 2018; Li et al., 2016; Shi, 2019) que se suman a las tradicionales formas de obtenerlos, como



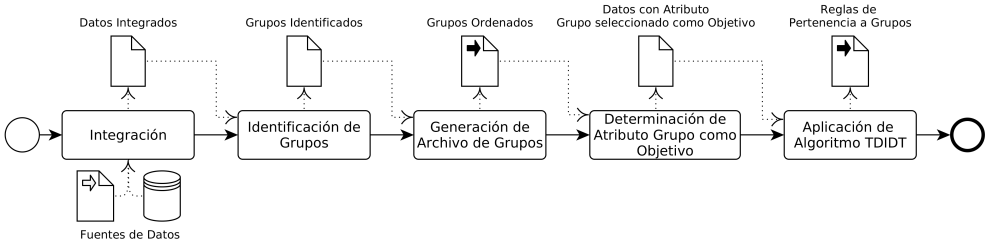
(a)



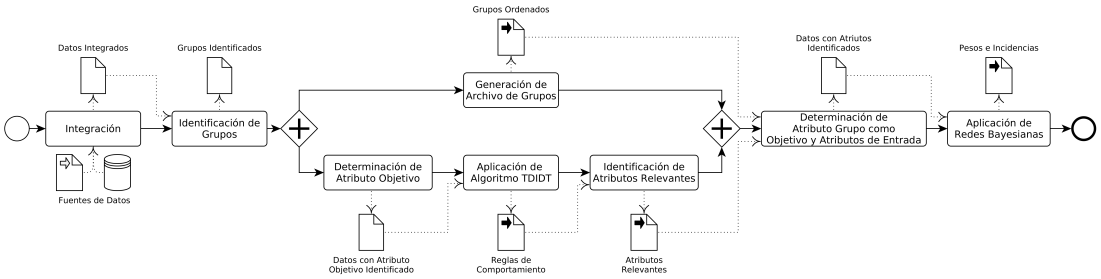
(b)



(c)



(d)



(e)

Figura 2.6: Procesos de explotación de información basados en sistemas inteligentes propuestos en (Britos, 2008; García-Martínez et al., 2013).

son la exploración y mapeo de campo, censos, análisis de terrenos, digitalización de mapas, o incluso simulaciones. La alta disponibilidad de este tipo particular de datos nos brinda oportunidades de descubrir nuevo conocimiento interesante y útil a partir de patrones no triviales implícitos en los mismos.

No obstante, las bases de datos espaciales poseen ciertas particularidades que las diferencian de las bases de datos tradicionales (relacionales o transaccionales). En primer lugar, una base de datos espaciales se define como sigue (Aggarwal, 2015; Manjula and Narsimha, 2014):

Definición 2.1 (Base de datos espaciales) *Una base de datos espaciales es un conjunto de tuplas (u objetos espaciales)* $D = \{T_1, T_2, \cdots, T_d\}$*, donde cada tupla*

$$T_i = (S_i^1, S_i^2, \cdots, S_i^m, X_i^1, X_i^2, \cdots, , X_i^n)$$

es un objeto espacial conformado por m atributos espaciales S_i^j *y n atributos no espaciales* X_i^k*.*

Los atributos no espaciales (o de comportamiento) son dimensiones que almacenan valores relacionados con mediciones realizadas en un lugar o extensión del espacio, mientras que los atributos espaciales, también llamados contextuales, detallan dicha ubicación o extensión y dan un contexto a los valores de los atributos no espaciales.

Póngase como ejemplo el caso de estaciones meteorológicas ubicadas en la extensión del país. Cada una de ellas realiza mediciones periódicas sobre distintos fenómenos atmosféricos que son almacenados en atributos no espaciales como temperatura, humedad, velocidad del viento, entre otros, mientras que la ubicación de dicha medición se registra a través de coordenadas en dos atributos espaciales: latitud y longitud.

Existen dos modelos de representación de datos espaciales computacionalmente: la representación vectorial y la representación mediante grillas (*raster*). Ambas estrategias permiten modelar distintas características del espacio, por lo que optar entre una u otra depende de las necesidades del dominio del problema.

A continuación se describen brevemente las particularidades de cada tipo de modelado en las secciones 2.4.1 y 2.4.2 y se brinda un método de traducción entre ambos en la sección 2.4.3.

2.4.1. Datos espaciales vectoriales

Una de las formas de modelar situaciones u objetos relacionados con un lugar determinado del espacio es la utilización del denominado modelo vectorial, consistente en conjuntos de atributos espaciales y no espaciales por cada entidad de la base de datos, tal como se ha definido en el apartado anterior.

Se utilizan tres modelos geométricos para representar estas entidades, dependiendo de sus características, expansión espacial y grado de granularidad requerido por el dominio del problema: los puntos, las líneas y los polígonos (Fig. 2.7) (Chang, 2006; Malerba, 2008; Manjula and Narsimha, 2014) .

En primer lugar, los puntos son utilizados para modelar objetos u eventos que ocurren en ubicaciones puntuales del espacio. Ejemplos factibles de ser modelados mediante puntos son centrales meteorológicas, accidentes automovilísticos, crímenes urbanos, intersecciones de carreteras, lugares de avistamiento de aves, entre innumerables otros ejemplos.

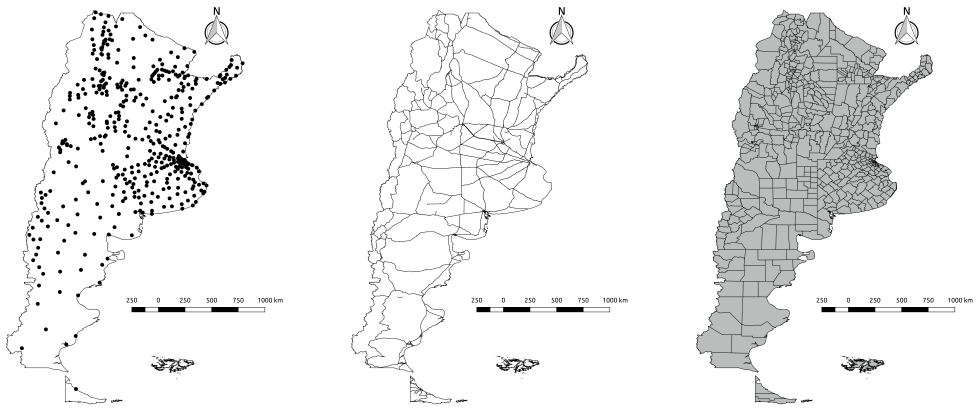
Los puntos son codificados, generalmente, mediante coordenadas tales como ordenadas y abscisas en un plano cartesiano de dos dimensiones, latitud y longitud en una base de datos geográfica, o incluso mediante coordenadas polares.

Seguidamente, las líneas son utilizadas para representar extensiones lineales en el espacio. Estas son herramientas útiles para modelar carreteras, ríos, recorridos de transporte público, tendidos de redes eléctricas, tuberías de gas, trayectorias de animales, etcétera.

Su codificación se realiza usualmente a través de una secuencia de puntos donde el primero se corresponde con el punto de origen de la línea, los consecuentes sus puntos de inflexión en el espacio, y por último el lugar de su finalización. Cuantos más puntos posea la secuencia, mayor será la precisión del recorrido de la misma.

Este tipo de objetos espaciales, en contraste con los puntos, poseen una propiedad que se adiciona a la ubicación del objeto, su longitud, definida generalmente como la suma de las distancias de los segmentos que conforman la misma.

Por último, los polígonos se utilizan para modelar extensiones bidimensionales del espacio como regiones o áreas. Parcelas de terreno, zonas de probabilidad de ocurrencia de eventos y zonas urbanas son sólo algunos ejemplos de situaciones factibles de ser representadas mediante polígonos.



(a) (b) (c)

Figura 2.7: Ejemplo de utilización de los distintos tipos de datos espaciales vectoriales. (a) Puntos representando ciudades cabeceras de departamentos de Argentina; (b) Líneas representando rutas nacionales de Argentina; (c) Polígonos representando departamentos de Argentina. Datos proporcionados por el Instituto Geográfico Nacional de la República Argentina.

Su expresión en las bases de datos se realiza mediante una línea o conjunto de líneas cerradas, por lo que a su vez una sucesión de puntos cuyo primer y último elemento es el mismo puede codificarlas de forma equivalente. Adicionalmente, una propiedad extra es factible de ser extraída de los polígonos: su área.

En una base de datos espaciales comúnmente coexisten objetos espaciales de tipos, representaciones y conjuntos de atributos distintos. Por caso, ciudades en un mapa pueden ser representadas mediante polígonos con atributos sobre sus características demográficas, mientras que determinadas ubicaciones de interés dentro de cada ciudad pueden ser modeladas utilizando puntos y atributos sobre los horarios y costos para visitarlos. Conjuntamente, las redes de calles pueden representarse mediante líneas y atributos sobre las condiciones del tráfico o el estado de la ruta.

Optar por un tipo de objeto espacialmente referenciado depende además de la resolución a la que se trabaje y del propósito de la base de datos. Ciudades pueden ser representadas como un simple punto, si estamos considerando una baja resolución y sólo necesitamos la ubicación de las mismas, o mediante áreas, si nos interesa su extensión en el espacio.

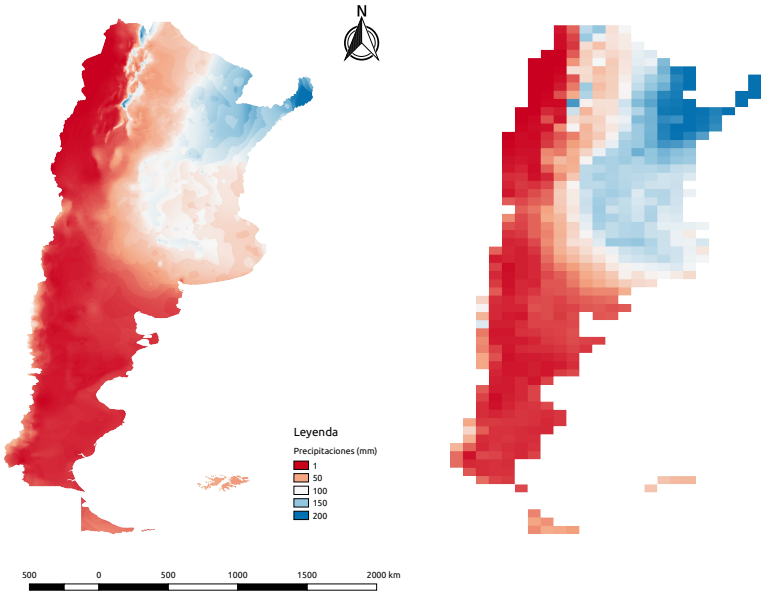


Figura 2.8: Precipitación media mensual de noviembre en la República Argentina (IDESA, 2017). La figura de la izquierda muestra un objeto *raster* de alta resolución, mientras que la imagen de la derecha muestra una resolución menor.

2.4.2. Datos espaciales basados en grillas

Si bien los datos vectoriales son una excelente alternativa para representar objetos espaciales discretos no resultan adecuados para modelar fenómenos que varían de forma continua por el espacio como la temperatura o las precipitaciones, por ejemplo. Ante estos casos, entonces, se utilizan los denominados datos *raster* o basados en grillas.

Este tipo de datos se modela a partir de una discretización del espacio subyacente a través de una grilla donde cada una de las celdas que la conforman almacena información agregada sobre la porción del espacio en cuestión. Por lo general, este tipo de datos son utilizados para codificar información obtenida mediante métodos de censado remoto, tales como imágenes satelitales o fotografías aéreas.

La capacidad de representación de la grilla está dada por su resolución: cuanta mayor sea la misma se podrá almacenar información de porciones más pequeñas del espacio, haciendo más suave la transición entre los valores de las celdas vecinas pero necesitando mayor poder de procesamiento. (Fig. 2.8).

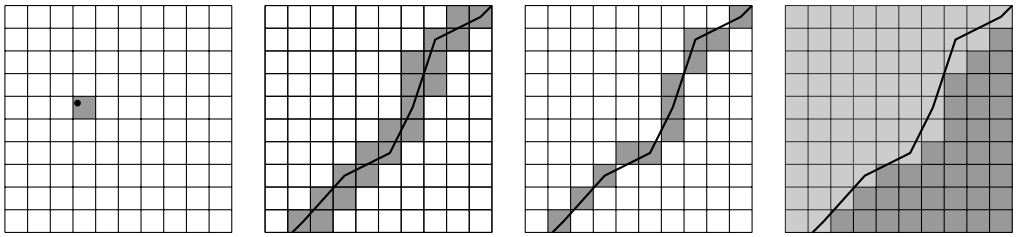


Figura 2.9: Representación en una grilla de objetos vectoriales. (a) Puntos; (b) Líneas gruesas; (c) Líneas delgadas; (d) Polígonos.

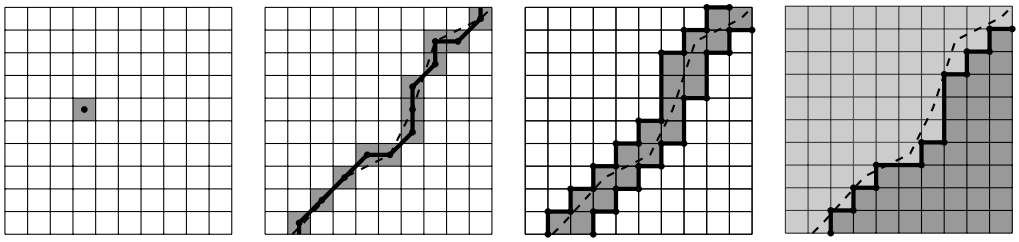


Figura 2.10: Generación de datos vectoriales a partir de una grilla. Objeto vectorial original representado mediante una línea a trozos. (a) Puntos; (b) Líneas delgadas; (c) Líneas gruesas; (d) Polígonos.

2.4.3. Traducción de representación

Si bien los datos raster modelan un espacio continuo, puntos, líneas y polígonos pueden ser representados mediante la selección de celdas de la grilla. Los puntos son modelados como un único píxel, mientras que las líneas y las regiones consisten en un agrupamiento de píxeles vecinos con extensión lineal o bidimensional, según corresponda.

El proceso de representar objetos vectoriales discretos en un espacio continuo se denomina rasterización (*rasterization*), mientras que el proceso de traducir los objetos raster como objetos vectoriales se denomina vectorización (*vectorization*).

En el primer caso, la estrategia a utilizar consiste en superponer los objetos espaciales discretos sobre el espacio mallado y asignar a cada celda intersecada por los mismos los atributos no espaciales que correspondan. Si el objeto puntual es un punto se seleccionará la celda que lo contiene (Fig. 2.9a). Si más de un punto se localiza dentro de la misma celda se deberá evaluar la posibilidad de aumentar la resolución de la grilla o de agregar los datos para describir su contenido.

Para las líneas vectoriales existen dos estrategias principales: descripción de líneas delgadas y gruesas. En el último caso se asignará la información a todas las celdas del espacio que posean intersección con la línea, por más pequeño que sea el contacto (Fig.2.9b). La estrategia delgada, por el contrario, identifica la menor cantidad de celdas posibles para caracterizar el objeto espacial, mejorando la visualización pero reduciendo la cantidad de información disponible sobre el mismo (Fig. 2.9c).

Finalmente, en el caso de los polígonos se busca seleccionar mediante la utilización de los bordes gruesos mencionados anteriormente las celdas que corresponden a la región en cuestión (Fig. 2.9d).

La transformación inversa es posible mas no sin pérdida de información. En general se utilizan los centroides de las celdas para generar las coordenadas que conformarán los objetos vectoriales, sin embargo, las mismas pueden no corresponder con la posición exacta del punto modelado, cuestión que puede verse contrastada entre las figuras 2.9a y 2.10a.

Por otro lado, el tipo de estrategia utilizada para representar las líneas en la grilla influirá en la forma del vector generado. Una estrategia de línea gruesa formará una región de la cual, a partir de su frontera, pueden extraerse dos líneas entre las que se deberá optar (Fig. 2.10c); por el contrario, para una estrategia delgada se utilizarán los centroides de las celdas como puntos de inflexión (Fig. 2.10b). Últimamente, las regiones estarán definidas por las líneas frontera, tal como se realiza para la estrategia gruesa mencionada anteriormente.

Seleccionar el tipo y características apropiadas del modelo es crítico para la posterior extracción de relaciones implícitas de los datos espaciales (Malerba, 2008).

2.5. Relaciones espaciales

La dependencia espacial de los valores almacenados en la base de datos permite que distintas relaciones, generalmente binarias, puedan ser obtenidas entre los objetos espaciales (Malerba, 2008).

Definición 2.2 (Relación espacial) *Dada una base de datos espaciales D, una relación $R \subseteq D^2$ es llamada espacial si y sólo si está definida mediante un predicado binario $P(x, y)|x \in D \wedge y \in D$ que involucra a los atributos espaciales de los objetos evaluados.*

Tabla 2.7: Ejemplo de funciones de distancia.

Nombre	Fórmula
Distancia Euclidiana	$d_E(\vec{P},\vec{Q}) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$
Distancia de Chebyshov	$d_C(\vec{P},\vec{Q}) = max_{i=1}^{n}(\lvert p_i - q_i\rvert)$
Distancia Manhattan	$d_M(\vec{P},\vec{Q}) = \sum_{i=1}^{n}\lvert p_i - q_i\rvert$
Distancia Geográfica	$d_G(P,Q) = R\sqrt{(\Delta\phi)^2 + (cos(\phi_m)\Delta\lambda)^2}$

A modo de ilustración se puede definir la relación de vecindad entre dos objetos espaciales como sigue:

Definición 2.3 (Relación de vecindad entre datos espaciales bidimensionales) *Sea D una base de datos espaciales, la relación de vecindad $V \subseteq D^2$ está definida por todos los pares ordenados $(x,y) \in D^2 | [(x_{lat} - y_{lat})^2 + (x_{long} - y_{long})^2]^{\frac{1}{2}} \leq \delta$, $\delta \in \mathbb{R}^+$*

El cómputo de estas relaciones implica una operación *join* entre dos subconjuntos de datos haciendo que su cálculo sea, por lo general, costoso, ya que escala rápidamente en función de la cardinalidad de los mismos.

Las relaciones espaciales pueden clasificarse según su naturaleza en geométricas, cuando están basadas en los principios de la geometría euclidiana; direccionales, cuando se contempla su ubicación relativa en el espacio; topológicas, cuando las mismas no se ven afectadas por transformaciones como la traslación, rotación y escalado; e Híbridas, cuando intervienen más de una de las clases anteriormente enumeradas (Appice et al., 2003; Malerba, 2008). En las subsecciones siguientes se profundiza en cada una de estas clases.

2.5.1. Relaciones geométricas

Las relaciones geométricas entre dos objetos espaciales son aquellas que se basan en los principios de la geometría euclidiana.

El cálculo de las mismas involucra una función de distancia entre los elementos que participan de la relación, tales como la misma distancia euclidiana, la distancia de

$$DE-9IM(a,b) = \begin{pmatrix} dim(I(a) \cap I(b)) & dim(I(a) \cap B(b)) & dim(I(a) \cap E(b)) \\ dim(B(a) \cap I(b)) & dim(B(a) \cap B(b)) & dim(B(a) \cap E(b)) \\ dim(E(a) \cap I(b)) & dim(E(a) \cap B(b)) & dim(E(a) \cap E(b)) \end{pmatrix}$$

Figura 2.11: Modelo de 9 intersecciones de (Clementini et al., 1993, 1994).

Manhattan, la distancia de Chebyshov, la distancia geográfica, entre otras (Tabla 2.7), utilizando los atributos espaciales de los datos (Maiti and Subramanyam, 2018).

Este cálculo, además, depende de los tipos de geometrías que se ven involucradas (puntos, líneas o polígonos), considerando generalmente la menor distancia entre ambos elementos siendo este valor nulo en caso de que exista una relación de intersección o inclusión entre los objetos).

2.5.2. Relaciones topológicas

Las relaciones topológicas son independientes de los conceptos de distancia y dirección y no se ven afectadas por transformaciones lineales como traslación, escalado y rotación. Son ejemplos de estas propiedades, la intersección de dos objetos espaciales, la inclusión, la adyacencia, entre otras.

Generalmente, se utiliza el Modelo de 9 Intersecciones Dimensionalmente Extendido (Dimensionally Extended nine-Intersection Model, DE-9IM) propuesto por Clementini et al. (1993, 1994) para modelar de una manera estandarizada las relaciones para su uso en el análisis computacional.

DE-9IM consiste en una matriz de dimensión 3×3 que permite codificar las relaciones entre dos objetos bidimensionales.

El modelo distingue 3 partes de un objeto espacial: su interior ($I(a)$), su exterior($E(a)$) y su frontera ($B(a)$). Una relación topológica entre dos objetos espaciales puede ser descripta en términos de la dimensión de las intersecciones entre sus partes, expresadas tal como se muestra en la figura 2.11.

Esta matriz es utilizada para definir predicados binarios como “Igual a”, “Disjuntos”, “Toca a”, “Contiene a”, “Cubre a”, “Interseca con”, “Dentro de”, “Cubierto por”, “Cruza a” y “Se superpone a”.

Es importante tener en cuenta que algunas de estas interacciones pueden ser ab-

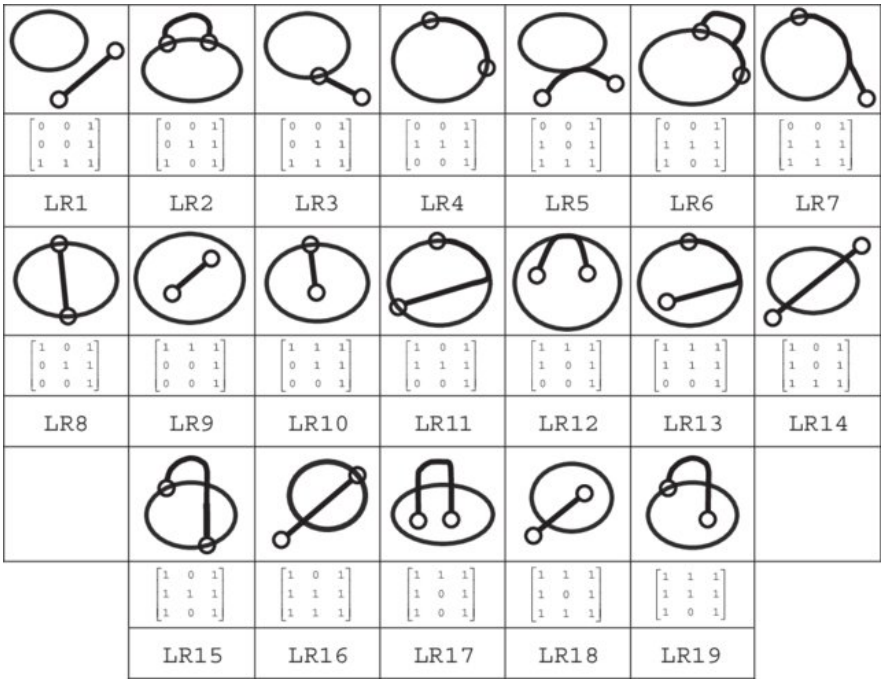


Figura 2.12: Relaciones topológicas significativas entre una región y una línea y sus correspondientes matrices representativas. Imagen extraída de (Du et al., 2010).

surdas de ser modeladas, de acuerdo a los tipos de objetos espaciales que intervienen. Por ejemplo, no resulta adecuado evaluar si un punto contiene a una línea. En la figura 2.12 podemos observar las posibles combinaciones de valores de la matriz DE-9IM que representan relaciones significativas entre una región y una línea.

2.5.3. Relaciones direccionales

Las relaciones direccionales involucran información sobre la orientación relativa a un objeto en el espacio.

Las métricas más comunes adoptadas para definir vecindarios son las ya mencionadas distancia Manhattan, también llamada “distancia de ciudad”, y la distancia de Chebyshov , llamada también “distancia del tablero de ajedrez”: los vecinos de un objeto dado son aquellos objetos que tienen distancia unitaria desde el punto de origen. El vecindario de un objeto, entonces, está constituido por los vecinos de dicho objeto y el objeto mismo.

Cuando se utiliza el modelo de Manhattan se obtienen cuatro vecinos (Figura 2.13, a) y cuatro direcciones principales a considerar al explorar el vecindario de un objeto

(a) (b)

Figura 2.13: Cuatro direcciones relativas a un punto en un espacio de dos dimensiones definidas utilizando (a) la distancia de Manhattan,y (b) la distancia de Chebyshov.

espacial: norte, sur, este y oeste; mientras que al utilizar el modelo de tablero de ajedrez son 8 las posibles direcciones (Figura 2.13, b): norte, noreste, este, sudeste, sur, sudoeste, oeste, noroeste.

Esta distancia unitaria puede variar dependiendo del dominio del problema. Si no es posible el cálculo de las distancias se puede utilizar el ángulo de incidencia entre ambos objetos espaciales para determinarlas.

2.5.4. Relaciones híbridas

Las propiedades híbridas surgen a partir de la combinación de propiedades clasificadas en dos o más categorías de relaciones diferentes. Por ejemplo, las relaciones de paralelismo y perpendicularidad entre dos líneas son tanto geométricas ya que se relacionan estrechamente con el espacio sobre el que son calculadas, como topológicas, ya que la propiedad se mantiene invariante frente a transformaciones de traslación, rotación y escalado.

Las relaciones híbridas no suelen ser calculadas específicamente como una categoría separada sino que se las incluye dentro de algunas de las categorías anteriormente listadas, según corresponda para cada relación.

2.6. Autocorrelación espacial

De acuerdo con la Primera Ley de la Geografía de Tobler, “Todo está relacionado con todo lo demás, pero las cosas cercanas se encuentran más relacionadas que las cosas lejanas”(Tobler, 1979).

Este enunciado nos menciona que los datos espaciales no son estadísticamente independientes, sino que aquellos que se encuentran en locaciones próximas del espacio tienden a parecerse, lo cual es consistente con las observaciones que se realizan en las áreas en las que su uso es habitual. Por ejemplo, las personas con características similares tales como ocupación y trasfondo tienden a vivir en los mismos vecindarios, los locales comerciales tienden a agruparse por rubro, entre otros.

Esta dependencia espacial es llamada autocorrelación espacial (*spatial autocorrelation*), y se define como la propiedad de variables aleatorias de tomar valores en ciertas posiciones del espacio separadas por una distancia determinada, que son más similares (autocorrelación positiva) o menos similares (autocorrelación negativa) que lo esperado para pares de ubicaciones asociadas de forma aleatoria (Cressie, 1992; Legendre, 1993).

Generalmente las técnicas de minería de datos asumen que los datos son independientes y distribuidos de forma idéntica, lo que provoca que las hipótesis o modelos sean inadecuados (Jiang and Shekhar, 2017).

2.7. Heterogeneidad espacial

La heterogeneidad espacial es otra de las características de los datos espacialmente referenciados que debe ser tenida en cuenta en la obtención de patrones de conocimiento.

Esta propiedad se refiere a que los objetos espaciales de la base de datos no siguen una distribución idéntica en todo el espacio. Un ejemplo puede darse cuando los objetos espaciales de la misma naturaleza pueden etiquetarse en distintas categorías en diferentes zonas del espacio (Miller and Han, 2009).

Por otro lado, la heterogeneidad espacial también se puede verificar en las tendencias que se dan entre las distintas variables de los objetos espaciales, como por ejemplo se observa en bases de datos de venta de inmuebles, donde puede ser posible que las casas antiguas posean un costo elevado en áreas rurales pero un costo considerablemente reducido en áreas urbanas. De esta forma, la antigüedad del inmueble no es un indicador directo del precio cuando todo el espacio de estudio es considerado .

Esta propiedad se denomina además Paradoja de Simpson Espacial y tiene como efecto que los modelos generados sobre la totalidad del espacio contemplado en la base

de datos no sean efectivos al aplicarse sobre regiones particulares (Jiang and Shekhar, 2017).

2.8. Minería de datos espaciales

La minería de datos espaciales es una extensión de la minería de datos tradicional que busca considerar las particularidades de este tipo de datos en la búsqueda de regularidades (Mennis and Guo, 2009), entre las que se encuentran las relaciones espaciales y las mencionadas autocorrelación espacial y heterogeneidad espacial (Atluri et al., 2018; Zeitouni, 2002).

Ejemplos de la incorporación de estas relaciones en el proceso de minería de datos se ve en procesos como el agrupamiento de datos espaciales, en el que se tienen en cuenta la distancia entre los objetos espaciales para la formación de los *clusters*, y en la búsqueda de asociaciones, donde las mismas deben incorporar predicados que involucren a los atributos espaciales de los datos evaluados.

Estas consideraciones extra no sólo que vuelven complejos a los patrones de conocimiento, sino que también complican el diseño de los métodos para obtenerlos, sin mencionar el aumento del costo computacional necesario para el cálculo de las relaciones espaciales y para la exploración del espacio de soluciones de los problemas de optimización asociados a las actividades de la minería de datos.

Según Li et al. (2016), existe una amplia variedad de patrones de conocimiento que generalmente no aparecen de forma aislada, sino que varios de estos tendrán que ser combinados para resolver problemas particulares. Entre estos patrones se encuentran los listados en la tabla 2.8.

En las subsecciones siguientes se mencionan las principales actividades de la minería de datos tradicional para obtener algunos de estos patrones y se mencionan las particularidades de su aplicación sobre datos espacialmente referenciados. sólo se ahonda en aquellas acotadas al alcance de esta tesis doctoral.

Tabla 2.8: Patrones de conocimiento sobre bases de datos espaciales. Adaptado y traducido de (Li et al., 2016).

Patrón	**Descripción**
Asociaciones	Asociación lógica entre diferentes conjuntos de entidades que relaciona uno o más objetos con otros, para el estudio de la frecuencia con la que dicha relación ocurre en la base de datos.
Patrón característico	Una característica común de un tipo de entidad o de múltiples tipos de entidades, para resumir atributos compartidos o similares de objetos.
Patrón discriminante	Una característica que distingue una entidad de otra, para comparar las particularidades de de tipos de objetos de interés.
Grupos	Colección de objetos de la base de datos que son similares entre sí y distintos a los objetos que no pertenecen al mismo, basado en cierto criterio a priori desconocido pero utilizando una medida de similitud previamente dada.
Patrón de clasificación	Criterio que determina con cierto nivel de confianza que un objeto de la base de datos pertenece a una determinada clase.
Patrón serial	Patrón que relaciona entidades o la dependencia funcional entre ciertos parámetros en un determinado lapso de tiempo, para analizar patrones espacio-temporales
Predicción	Determinación de valores para ciertos atributos de uno o varios tipos de datos desplazados en el tiempo o en el espacio a partir de los valores de las instancias de dichos tipos presentes en la base de datos.
Excepción	Anomalía u observación cuyo comportamiento desvía significativamente del comportamiento normal de la base de datos.

2.8.1. Grupos de datos espaciales

El agrupamiento (*clustering*) es una de las actividades de la minería de datos ampliamente utilizada. La misma permite determinar una partición $P = \{P_1, P_2, \cdots, P_k\}$ de elementos de las bases de datos de forma tal que los elementos que pertenecen a cada uno de los grupos de esta partición sean similares entre sí y diferentes a los elementos que forman parte de otros grupos (Han et al., 2011). Para esto se hace uso de algoritmos basados en estadísticas y aprendizaje automático como K-Means (Alg. 2.1) (MacQueen et al., 1967), SOM (Kohonen, 1982) y DBSCAN (Ester et al., 1996; Schubert et al., 2017).

Este tipo de algoritmos es utilizado en procesos de explotación de información como el proceso de descubrimiento de grupos, proceso de descubrimiento de reglas de pertenencia a grupos y proceso de ponderación de reglas de pertenencia a grupos, mencionados anteriormente en la sección 2.3, con el propósito de obtener particiones de elementos similares en el dominio del problema para tomar decisiones orientadas a cada uno de los grupos de datos en particular, según sus características.

Algoritmo 2.1 K-means.

Entrada: $X \leftarrow$ Conjunto de datos
Entrada: $k \leftarrow$ Número de grupos
Salida: Partición de X
  $C = c_1, c_2, \cdots, c_k \leftarrow$ generar k centroides aleatorios.
  repetir
    para todo $x_i \in X$ **hacer**
      asignar x_i al centroide $c_j \in C$ más cercano.
    fin para
    $C \leftarrow$ recalcular centroides
  hasta que el algoritmo converja
  devolver asignación de cada x_i a los centroides c_j

No obstante, al considerar la información espacial se diferencian tres tipos de agrupamiento distintos. En primer lugar el homónimo agrupamiento espacial calcula la similitud entre los objetos espaciales considerando los atributos espaciales de los mismos y relaciones espaciales entre ellos. Una lista extensa de algoritmos de agrupamiento espacial puede ser encontrada en (Han et al., 2001).

Segundo, el proceso de regionalización es un tipo particular de agrupamiento que busca formar grupos de objetos espaciales similares entre sí, con la restricción añadida de que los *clusters* deben ser espacialmente contiguos (regiones). Múltiples aplicacio-

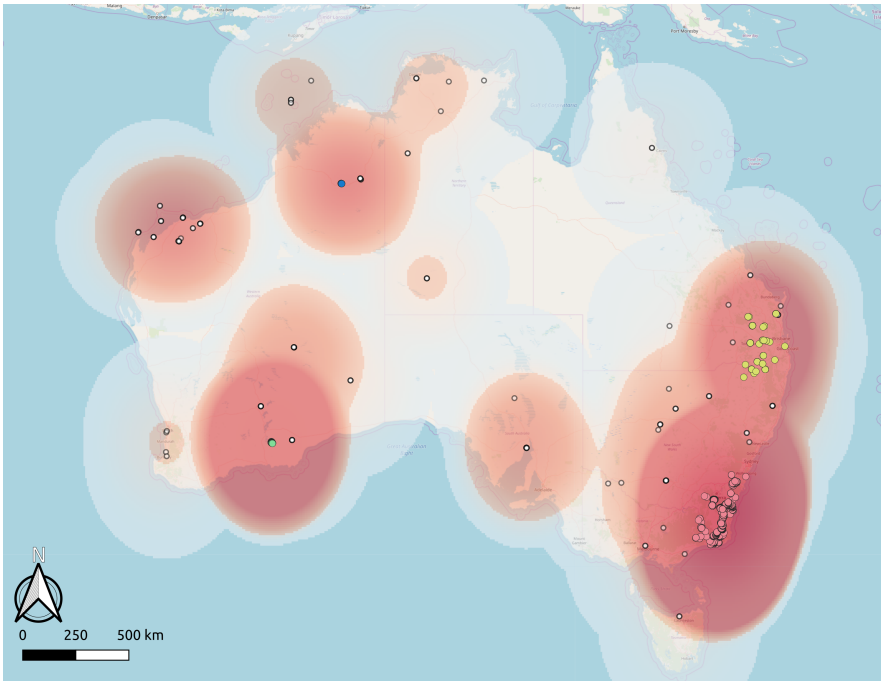


Figura 2.14: Ejemplo de puntos calientes según la densidad de focos de incendio en Australia el día 26 de enero de 2020. Datos extraídos del sistema FIRMS de NASA y procesados mediante el algoritmo DBSCAN.

nes en climatología, análisis de relieves, entre otros, requieren esta restricción adicional. Los métodos para regionalización pueden ser clasificados en tres tipos (Mennis and Guo, 2009): 1. *clustering* de variables no espaciales seguido por un reordenamiento de los grupos para cumplir con la restricción de contigüidad; 2. *clustering* con medida de similitud ponderada que contempla las propiedades espaciales como un factor de formación de los grupos; y 3. *clustering* con restricciones de contigüidad, que fuerza a los grupos a ser espacialmente contigüos durante el proceso de construcción de los mismos.

Por último, el análisis de patrones de puntos, también llamado análisis de regiones calientes (*hotspots*), se centra en la detección de concentraciones inusuales de objetos espaciales, tales como enfermedades, crímenes o accidentes de tráfico, en un determinado área (Fig. 2.14) (Brimicombe, 2007). Esta actividad suele realizarse además con herramientas estadísticas como la I de Moran que mide la autocorrelación espacial de los objetos espaciales considerados, determinando no sólo zonas calientes sino también zonas frías. Es importante recalcar que ambos criterios son distintos si bien están relacionados: mientras que el primero se centra en la densidad para generar los grupos el segundo calcula el índice de autocorrelación, mas es normal la existencia de *clusters*

densos de puntos con una alta correlación espacial (Li et al., 2007).

2.8.2. Asociaciones espaciales

El descubrimiento de asociaciones, primeramente introducido sobre bases de datos transaccionales tales como aquellas que registran compras de supermercados (Agrawal et al., 1993), busca encontrar subconjuntos de *items* (como tipos de productos), que se encuentran frecuentemente juntos en una transacción.

Definición 2.4 (Patrón de asociación frecuente) *Sea* $I = \{i_1, i_2, \cdots, i_m\}$ *un conjunto de items y* D *un conjunto de transacciones se dice que* $T \subseteq I$, $X \subseteq I$ *es un patrón de asociación frecuente si*

$$\frac{|\{T \in D | X \subseteq T\}|}{|D|} \geq s$$

siendo s *el soporte mínimo del patrón de asociación.*

Asimismo, es posible expresar dicho patrón como una regla de asociación $P \rightarrow Q$ con $P \cup Q = X$, haciendo uso de otra métrica denominada "confianza" en la que interviene la probabilidad condicional de que ocurra un subconjunto de instancias del patrón dado otro subconjunto de instancias del mismo (Aggarwal, 2015).

De forma equivalente, un patrón de asociación puede verse como un predicado compuesto $P_1 \wedge P_2 \cdots \wedge P_m$ sobre D, pudiendo calcular el soporte del mismo como la razón entre el cardinal del conjunto de verdad del predicado y la cardinalidad de la base de datos.

A partir de esta definición, un patrón de asociación espacial se define como un predicado compuesto donde por lo menos uno de los predicados atómicos que lo componen involucra a los atributos espaciales de los datos. Estos predicados espaciales pueden definir relaciones espaciales como las ya mencionadas con anterioridad en la sección 2.5.

Un ejemplo de patrón de asociación se puede enunciar como: "las estaciones de gas se encuentran cerca de la intersección de dos calles", pudiéndose modelar mediante predicados de la siguiente forma:

$$Es\_Estacion\_Gas(x) \wedge Es\_Calle(y) \wedge Es\_Calle(z) \wedge$$
$$Interseca(y, z) \wedge Cerca(x, y) \wedge Cerca(x, z)$$

Este ejemplo ilustra el uso de múltiples tipos de datos espaciales vectoriales: puntos para las estaciones de gas y líneas para las calles, y el uso de relaciones espaciales de tipos variados: relaciones topológicas al considerar la intersección entre calles y relaciones euclidianas por la relación de cercanía.

Debido a la gran cantidad de relaciones posibles entre datos espaciales, diferencia notable entre las bases de datos de este tipo y las bases de datos transaccionales, el costo computacional de la búsqueda de este tipo de regularidad se eleva considerablemente.

Otro problema es la enorme cantidad de posibles reglas que pueden ser generadas siendo muchas de ellas obvias o de conocimiento general, por lo que es necesario filtrarlas y enfocarse en aquellas que resulten interesantes para la toma de decisiones.

Entre los métodos para el descubrimiento de reglas de asociación espaciales se destacan los trabajos de (Appice et al., 2003; Bai et al., 2014; Chen et al., 2008; Cunjin and Xiaohan, 2017; Ladner et al., 2003; Lee, 2004; Malerba, 2008; Sha and Li, 2010; Sutjipto et al., 2017; Yang et al., 2005), entre muchos otros.

Por otro lado, un tipo particular de reglas de asociación espacial son los ampliamente estudiados patrones de co-localización.

Dado un conjunto de datos espaciales con propiedades categóricas predefinidas, un patrón de co-localización constituye un subconjunto de estas categorías que se localizan frecuentemente en un mismo vecindario. En este escenario, todas las instancias del patrón deben ser vecinas entre sí, conformando un clique (Shekhar and Huang, 2001).

Esta clase de patrones permiten modelar especies de animales con relación de mutualismo o simbiosis, tipos de crímenes asociadas a características urbanísticas, entre otros ejemplos.

Si bien la cantidad de relaciones a ser extraídas para la detección de esta clase de regularidad es menor que en la búsqueda de asociaciones al considerarse sólo la relación euclidiana de vecindad, la restricción adicional de cumplimiento simultáneo de la misma entre todos los elementos de la instancia supone una complejidad extra.

En este área de estudio destacan los aportes en materia algorítmica de Celik et al.

(2007); Shekhar and Huang (2001); Wang et al. (2008); Yoo et al. (2005); Yu (2016), entre varios otros.

2.8.3. Descubrimiento de anomalías espaciales

Una anomalía se define de manera informal como observaciones que parecen ser inconsistentes con el resto de los datos de la base de datos, o que desvían de las demás observaciones lo suficiente como para sospechar que fueron generados por un mecanismo distinto (Shekhar et al., 2011).

Las anomalías suelen clasificarse en tres tipos distintos, según el contexto en el cual son relevantes: en primer lugar, las anomalías globales son aquellas observaciones o datos que difieren de la totalidad de la base de datos remanente; luego, las anomalías contextuales o locales poseen valores anómalos en comparación a un vecindario de los datos, como es el caso de un día caluroso en el invierno, cuyas temperaturas no serían extrañas al considerar todo el año; por último, las anomalías colectivas corresponden a un conjunto de datos que constituyen una anomalía por el hecho de encontrarse juntas, como suele darse en un electrocardiograma al variar el ritmo de una progresión de valores.

Al tratarse de datos espacialmente referenciados la definición de anomalía se altera sutilmente:

Definición 2.5 (Anomalía espacial) *una anomalía espacial se define como una observación del conjunto de datos cuyo valor de sus atributos no espaciales difiere de los valores de los atributos no espaciales de los objetos que se encuentran en un vecindario del primero (Bansal et al., 2016; Deepak, 2016)*

Informalmente, entonces, una anomalía espacial es una inestabilidad local (en valor de sus atributos no espaciales), por lo que corresponden con anomalías locales o contextuales. A modo de ejemplo, dada una base de datos de edificios en una ciudad, una casa de 5 años de antigüedad constituye una anomalía en un barrio histórico, mientras que en un barrio residencial relativamente nuevo no lo es (Rottoli et al., 2018; Shekhar et al., 2011).

La definición de estos vecindarios de búsqueda constituye un desafío, siendo necesaria información del dominio del problema para establecer un perímetro de búsqueda. Asimismo, el valor de un atributo no espacial puede variar gradualmente en el espa-

cio sin que esto constituya una anormalidad: siguiendo con el ejemplo anterior, en un área urbana la edad de los edificios varía gradualmente a medida que nos alejamos del centro histórico.

Si bien, la detección de outliers globales es realizada en general mediante técnicas que se encuentran bien estudiadas y documentadas en la actualidad, como las distancias de Mahalanobis que hacen uso de estimaciones robustas de covarianzas, para la búsqueda de anomalías locales es necesario tener en cuenta las consideraciones anteriormente mencionadas.

En este contexto, Ernst and Haesbroeck (2017), Schubert et al. (2014) y Cai et al. (2013) aportan excelentes revisiones sobre el estado del arte en cuanto a técnicas algorítmicas para la búsqueda de anomalías espaciales, por lo que se recomienda ahondar en esta literatura para obtener mayores detalles al respecto.

CAPÍTULO 3

DESCRIPCIÓN DEL PROBLEMA

> "Do you know why I stopped being Delight, my brother? I do. There are things not in your book. There are paths outside this garden."
>
> *Neil Gaiman, Brief Lives*

En la extensión de esta sección se describe la problemática que se desprende del análisis del estado del arte sobre el marco teórico desarrollado en el capítulo anterior. Esta problemática se detalla a continuación en la sección 3.1, mientras que en la sección 3.2 se delinean los objetivos de la investigación (sección 3.2.1) y las preguntas que guían el desarrollo de la misma (sección 3.2.2).

3.1. Estado del arte y problemática detectada

En el capítulo anterior se ha descripto cómo las distintas actividades de minería de datos deben ser adaptadas para considerar la dimensión espacial de las bases de datos implicando la extracción de relaciones espaciales euclidianas, direccionales, topológicas o mixtas, según el dominio del problema, y considerando fenómenos como la heterogeneidad espacial y la autocorrelación espacial.

Por cada una de estas actividades se han desarrollado cientos de algoritmos distintos, en su mayoría diseñados con dos posibles propósitos, a veces simultáneos: incrementar la eficiencia o solucionar problemas que se evidencian en escenarios específicos al utilizar ciertos datos particulares.

En cuanto a agrupamiento de datos espaciales, por mencionar algunos ejemplos, Ng and Han (1994) proponen un algoritmo para agrupamiento denominado CLARANS y utilizan este método para el minado de atributos espaciales por un lado y atributos no espaciales por otro utilizando una herramienta de aprendizaje distinta. Estos dos métodos se denominan SD(CLARANS) y NSD(CLARANS). Asimismo, Tung et al. (2001) también diseña un algoritmo basado en CLARANS, denominado COD-CLARANS, que permite la obtención de clusters considerando obstáculos entre los objetos espaciales. Algo similar ocurre en la propuesta de Sagvekar et al. (2013), que modifican este algoritmo base para agrupar polígonos. Además, Wei et al. (2009) proponen una variante de DBSCAN denominada LD-BSCA que permite considerar *clusters* de densidad variable con un número mínimo de parámetros de entrada. Wang and Wang (2012), por otro lado, desarrollan métodos para agrupar "puntos con dirección", una forma de representar objetos en el contexto de trayectorias pero sin considerar el aspecto espacial. Kim and Cho (2019) desarrollan un algoritmo de *clustering* que mejora la diferenciación de grupos basados en densidad cuando estos están muy próximos entre sí mejorando también la tolerancia al ruido. (Zhang et al., 2007) propone un acercamiento mediante la penalización de los resultados del algoritmo de agrupamiento para que los *clusters* obtenidos incorporando atributos no espaciales se mantengan separados, sin sobreponerse unos con otros.

Por otro lado, en lo que concierne al descubrimiento de asociaciones, algunos trabajos que ejemplifican lo anteriormente expuesto son los trabajo de Appice et al. (2003) y de Lisi and Malerba (2002), que permiten una alta capacidad de representación de relaciones entre objetos espaciales mediante programación lógica inductiva. También el trabajo de An-yang and Wei-dong (2005) que permite la consideración de múltiples medidas de soporte mínimo. Chen et al. (2008) propone un método que considera la autocorrelación espacial a través de una estructura celular. Bai et al. (2014) usa un modelo de extracción de reglas de asociación basado en teoría de conjuntos rígidos y difusos. Ladner et al. (2003) combina técnicas de minería de datos espacial y minería de datos difusa para lidiar con la incertidumbre que se encuentra en las bases de datos espaciales. Sha and Li (2010) propone un algoritmo para detección de patrones locales considerando la heterogeneidad espacial utilizando un acercamiento basado en densidad para evaluar el soporte de las asociaciones en áreas determinadas.Cunjin and Xiaohan (2017) presenta un algoritmo diseñado para descubrir relaciones espaciales relacionadas con fenómenos climatológicos como el fenómeno del Niño y la Niña.Yang et al. (2005) introduce el uso de grafos para modelar las relaciones entre

objetos espaciales.

En relación al descubrimiento de anomalías en bases de datos espaciales, algunos ejemplos de propuestas de la bibliografía con este objetivo se ven en trabajos como el de Lu et al. (2004), que incorpora múltiples dimensiones no espaciales a la búsqueda, en los métodos basados en algoritmos de agrupamiento de Cai et al. (2013) para la búsqueda de patrones en bases de datos multivariadas teniendo en cuenta el fenómeno de la autocorrelación espacial, en el trabajo de Shi et al. (2016) que propone un proceso basado en triangulación de Delaunay, o Xu et al. (2017) que propone un algoritmo adaptativo no parametrizado y lo aplica sobre redes de sensores inalámbricos, o incluso Shi et al. (2018) que propone un método basado en grafos para la búsqueda de anomalías relacionadas con la co-localización de eventos espaciales.

Asimismo, se evidencia una amplia variedad de trabajos en el estado del arte que utilizan algoritmos de minería de datos tradicionales para analizar datos espaciales sin considerar las cuestiones previamente mencionadas. En el caso del agrupamiento de datos espaciales, por ejemplo, la propuesta de Estivill-Castro and Houle (2001) permite agrupar datos en el espacio por su vecindad o densidad, considerando *outliers*, pero no permite incorporar de forma separada las dimensiones no espaciales de los datos. De esta forma los datos no son manipulados aprovechando adecuadamente las posibilidades que brinda la contextualización de los mismos.

Por otro lado, la amplia variedad de opciones algorítmicas generalmente no se encuentran implementadas en las herramientas tecnológicas para que su aplicación pueda ser realizada de forma inmediata. Esto interfiriere en la selección de alternativas, que generalmente han sido diseñadas para abordar problemas particulares tal como se ha elaborado anteriormente. Se suma además el hecho de que esta actividad de selección y diseño usualmente se realiza de forma artesanal mediante prueba y error, lo cual hace que se torne altamente complejo (Tripathy and Panda, 2017).

Además, en algunos trabajos del estado del arte se ignoran aspectos relacionados con la contextualización de la información espacial al diseñar las soluciones utilizando algoritmos no diseñados para tal fin. Por ejemplo, TarakaSravani et al. (2017) proponen un enfoque consistente en la aplicación de métodos de agrupamiento tradicional en distintas localidades espaciales y la integración de cada agrupamiento en jerarquías espaciales superiores. Varghese et al. (2013) realizan una enumeración de algoritmos de agrupamiento considerados como espaciales, mas esta lista de algoritmos incluye opciones tradicionales como el ya mencionado K-means, que no permiten considerar

relaciones espaciales complejas ni diferenciar entre atributos espaciales y no espaciales (si bien es cierto que algunas de estas opciones mencionadas en este trabajo se pueden utilizar en el contexto de datos espaciales, es necesaria una adecuada selección de parámetros para que los resultados sean útiles para la inteligencia de negocio). Babu et al. (2015) utiliza algoritmos de descubrimiento de asociaciones transaccionales a partir de transacciones generadas agrupando datos de cáncer por región geográfica. No se tienen en cuenta relaciones ni propiedades espaciales. Algo similar ocurre con el trabajo de Faridi et al. (2015), al utilizar el algoritmo apriori sin considerar ninguna propiedad ni relación espacial, cuestión que se evidencia además en múltiples trabajos revisados en (Gandhi and Armstrong, 2016), principalmente en el área de la agricultura. En contraste, Sadat et al. (2015), por ejemplo, realizan un excelente trabajo con respecto a este último tipo de regularidad, considerando relaciones y atributos difusos.

Con respecto al descubrimiento de anomalías espaciales, los métodos del estado del arte resultan apropiados pero de difícil acceso para los analistas, con algunas excepciones como las métricas LOF (que están pensadas para determinar anomalías locales pero no hacen una distinción explícita entre atributos espaciales y no espaciales). Tales son los casos de la propuesta presentada en (Zheng et al., 2017), aplicada para la determinación de anomalías espaciales relacionadas con *Hydraulic fracturing*, el trabajo de Van Zoest et al. (2018) centrado en evaluar extrañezas en la calidad del aire o la propuesta de (Singh and Lalitha, 2018).

Por estas razones mencionadas en los párrafos anteriores se plantea la necesidad del diseño de procesos de explotación de información que organicen la manipulación de datos espaciales considerando la extracción de relaciones espaciales, la búsqueda de patrones locales y la autocorrelación espacial para dar respuesta a problemas de negocio relacionados con este tipo particular de datos (Li et al., 2015). Esto se fundamenta en el planteo de (Li et al., 2016): resulta generalmente necesario combinar distintos patrones de conocimiento y estrategias para dar respuesta a problemas de negocio particulares.

Estos procesos deberán ser suficientemente flexibles para adaptarse a múltiples situaciones particulares mediante la consideración de múltiples tipos distintos de datos vectoriales y de relaciones espaciales entre los mismos.

Se buscará que los procesos puedan no solo extraer información útil para la toma de decisiones, sino que también sean implementables con herramientas de fácil acceso (Li et al., 2016), utilizando algoritmos y métodos que se encuentren disponibles en el

mercado. Esto posibilitará además la construcción de modelos prototipos y la selección de las mejores alternativas para cada conjunto de datos y problema particular.

3.2. Sumario de Investigación

Seguidamente se enuncian en la sección 3.2.1 el problema técnico de investigación y en la sección 3.2.2 las preguntas de investigación que se derivan del análisis de la situación planteada con anterioridad y dan lugar al trabajo doctoral materializado en la presente tesis.

3.2.1. Problema técnico de investigación

Se ha redactado el problema técnico de investigación (Wieringa, 2014) derivado del análisis de lo expuesto en la sección anterior de la siguiente manera (Tabla 3.1):

> Diseñar un conjunto de procesos de explotación de información para la extracción de conocimiento para la toma de decisiones estratégicas, a partir de conjuntos de datos espaciales vectoriales, considerando relaciones espaciales y las propiedades de autocorrelación y heterogeneidad espacial. Asimismo, deben ser factibles de ser implementados mediante el uso de tecnologías informáticas de fácil acceso.

Tabla 3.1: Definición del problema de diseño abordado en la tesis doctoral

Aspecto	**Descripción**
Artefacto	Conjunto de procesos de explotación de información.
Contexto	Datos espacialmente referenciados vectoriales de múltiples formatos (puntos, líneas, polígonos).
Propiedades	Consideración de propiedades de autocorrelación y heterogeneidad espacial. Implementables mediante el uso de tecnologías de fácil acceso. Facilidad de interpretación.
Propósito	Toma de decisiones estratégicas.

3.2.2. Preguntas de investigación

A partir de lo planteado en la sección anterior se plantean las siguientes preguntas de investigación que colaboran, mediante su respuesta, con el cumplimiento del mismo:

P1 ¿Cuáles características asociadas a los datos espacialmente referenciados deben considerarse en la extracción de cada tipo de patrón de conocimiento?

P2 ¿Cuáles son las actividades necesarias para la extracción de dicho conocimiento?

P3 ¿Cuáles son los formatos de datos apropiados para la entrada y la salida de las actividades de cada uno de los procesos?

P4 ¿De qué manera se pueden implementar estos procesos?

En los capítulos siguientes se expone el diseño de los artefactos que dan respuesta a los interrogantes mencionados y se desarrollan sus correspondientes validaciones.

CAPÍTULO 4

SOLUCIÓN

> "Dream the world. Not this pallid shadow of reality. Dream the world the way it truly is."
>
> *Neil Gaiman, A Dream of a Thousand Cats*

En el presente capítulo se abordará la solución del problema planteado previamente. Para esto, primeramente, en la sección 4.1 se presentan y detallan las herramientas teórico-conceptuales que han sido utilizadas para el diseño de los procesos de explotación de información, entre las que se incluyen la teoría de grafos (sección 4.1.1) y la teoría de conjuntos difusos (sección 4.1.2). Seguidamente, en la sección 4.2 se da lugar al detalle de los procesos de exploración de información diseñados como respuesta a los planteamientos realizados, incluyendo un proceso para el descubrimiento de grupos espaciales (sección 4.2.1), un proceso para el descubrimiento de anomalías espaciales (sección 4.2.2), un proceso para el descubrimiento de asociaciones espaciales (sección 4.2.3) y un proceso para el descubrimiento de patrones de co-localización (sección 4.2.4). Por último, para finalizar el capítulo, se plantea el modelado de relaciones espaciales mediante lógica difusa y su uso en dos de los procesos diseñados en la sección 4.3.

4.1. Herramientas preliminares

Las soluciones propuestas en el presente trabajo hacen uso de herramientas provenientes de la teoría de grafos y de la teoría de conjuntos difusos para modelar las

relaciones entre los datos espaciales. Consecuentemente resulta necesario introducir preliminarmente las bases conceptuales de estas teorías, siendo esto realizado en las siguientes subsecciones.

La sección 4.1.1 describe brevemente los principales conceptos de la teoría de grafos usando como base para ello las obras "*Discrete Mathematics and Its Applications*" de Rosen and Krithivasan (2012), 7ma. edición, y "Discrete Mathematics with Applications" de Epp (2011), 4ta. edición. Luego, en la sección 4.1.2 se describe brevemente la teoría de conjuntos difusos utilizando como base el libro titulado "Los conjuntos borrosos y su aplicación a la programación lineal" de Lazzari et al. (2001).

4.1.1. Teoría de grafos

En esta sección se describen los conceptos básicos relacionados con la teoría de grafos (Sección 4.1.1.1), clases de grafos con características particulares y útiles para los propósitos de los diseños presentados en esta tesis (Sección 4.1.1.2), los digrafos como herramientas relacionadas con los grafos (Sección 4.1.1.3), el uso de grafos y digrafos como modelos de situaciones reales (Sección 4.1.1.4) y una breve introducción al uso de herramientas de minería de datos sobre estas estructuras (Sección 4.1.1.5).

4.1.1.1. Grafos

Los grafos son estructuras consistentes en un conjunto finito de vértices y un conjunto finito de aristas. Cada arista posee una relación con uno o dos vértices a los que se denominan "puntos extremos" del mismo.

Formalmente, un grafo se define como sigue:

Definición 4.1 (Grafo) *Un grafo G es una tupla (V, E, φ), donde V es un conjunto finito no vacío de vértices, E es un conjunto finito de aristas, y $\varphi : E \to \{\{u, v\} \mid u \in V \wedge v \in V\}$ es la función de incidencia, que especifica qué par (sin orden) de vértices le corresponde a cada arista.*

Asimismo, los conceptos asociados a los grafos son los siguientes:

Definición 4.2 (Incidencia) *Dado un grafo $G = (V, E, \varphi)$, una arista $e \in E$ y un vértice $v \in V$ del grafo, se dice que e incide sobre $v \Leftrightarrow v \in \varphi(e)$*

Definición 4.3 (Adyacencia) *Dados dos vértices cualesquiera* v_1 *y* v_2 *de un grafo* $G = (V, E, \varphi)$*, se dice que* v_1 *es adyacente a* $v_2 \Leftrightarrow \exists e \in E, \varphi(e) = \{v_1, v_2\}$

Definición 4.4 (Aristas paralelas) *Dadas dos aristas cualesquiera* e_1 *y* e_2 *de un grafo* $G = (V, E, \varphi)$*, se dice que* e_1 *es paralela a* $e_2 \Leftrightarrow \varphi(e_1) = \varphi(e_2)$

Definición 4.5 (Bucle) *Dada una arista cualquiera* e *de un grafo* $G = (V, E, \varphi)$*,* e *es un bucle* $\Leftrightarrow |\varphi(e)| = 1$

Definición 4.6 (Vértice aislado) *Dado un vértice cualquiera* v *de un grafo* $G = (V, E, \varphi)$ v *es un vértice aislado* $\Leftrightarrow \forall e \in E, v \notin \varphi(e)$

Definición 4.7 (Grado de un vértice) *Dado un grafo* $G = (V, E, \varphi)$ *el grado de un vértice* $v \in V$ *se define como* $g(v) = |\{e \in E | v \in \varphi(e)\}|$

De forma coloquial, podemos definir la incidencia como una relación entre arista y vértice: se dice que una arista incide sobre los vértices que son sus puntos extremos. La adyacencia, por otro lado, es una relación entre vértice y vértice: se dice que dos vértices son adyacentes si existe alguna arista que incida sobre ambos. Asimismo, las aristas paralelas son aquellas que tienen los mismos puntos extremos, los bucles son aristas que inciden sobre un solo vértice, y los vértices aislados son vértices sobre los que no incide ninguna arista.

Estas estructuras permiten modelar relaciones entre distintos elementos de un escenario real o virtual, pudiéndose representar en un plano donde los vértices son puntos y las aristas son líneas que conectan los puntos sobre los que inciden (Fig. 4.1a).

4.1.1.2. Grafos particulares

Entre todos los grafos posibles de ser construidos se distinguen algunos que poseen características particulares, algunos de los cuales pueden ser herramientas de utilidad para la minería de datos espaciales.

En primer lugar, los grafos denominados grafos simples se caracterizan por no poseer bucles ni aristas paralelas. En este caso es posible prescindir de la función de incidencia pues solo existe como máximo un arista en E por cada par de vértices de V (Def 4.8).

Definición 4.8 (Grafo simple) *Un grafo* $G = (V, E, \varphi)$ *es un grafo simple sii* $\forall e \in E, |\varphi(e)| = 2 \wedge \forall e_1, e_2 \in E, \varphi(e_1) = \varphi(e_2) \rightarrow e_1 = e_2$

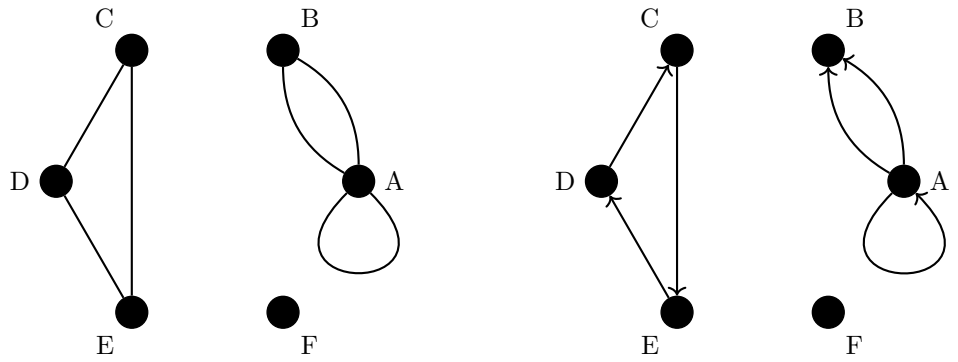


Figura 4.1: (a) Ejemplo de grafo con un vértice aislado etiquetado como F, aristas paralelas que inciden sobre los vértices A y B, y un bucle que incide sobre A. (b) Ejemplo de digrafo con vértice aislado etiquetado como F, aristas estrictamente paralelas que inciden sobre los vértices A y B y un bucle que incide sobre el vértice A.

Luego, los grafos completos son grafos simples donde existe un arista entre cada par de vértices. Esto es, un grafo simple cuyo número de aristas es máximo (4.9, fig. 4.2).

Definición 4.9 (Grafo completo) *Dado un grafo simple $G = (V, E, \varphi)$, es un grafo completo sii $\forall x, y \in V, \exists e \in E \mid \varphi(e) = \{x, y\}$. Este grafo se denota K_n, donde $n = |V|$ y se cumple que $|E| = \frac{n(n-1)}{2}$.*

Por otro lado, los grafos bipartitos completos, denominados $K_{n,m}$ son grafos simples cuyo conjunto de vértices puede partirse en dos subconjuntos, de tal forma que todos los vértices de cada subconjunto sean adyacentes a los vértices del otro subconjunto y no existan aristas entre los vértices de un mismo subconjunto (Def. 4.10).

Definición 4.10 (Grafo bipartito completo) *Dado un grafo simple $G = (V, E, \varphi)$, G es un grafo bipartito sii $\exists\{V_1, V_2\} \mid V_1 \cup V_2 = V \wedge V_1 \cap V_2 = \emptyset$ de tal forma que $\forall x \in V_1, \forall y \in V_2, \exists e \in E | \varphi(e) = \{x, y\}$ y además $\forall e \in E,\ \varphi(e) = \{x, y\} \rightarrow x \in V_1 \wedge y \in V_2$. Este grafo se denota como $K_{m,n}$, donde $|V_1| = m$, $|V_2| = n$ y $|E| = m * n$*

Un subgrafo de un grafo G es un grafo cuyo conjunto de vértices y conjuntos de aristas son subconjuntos del conjunto de vértices y aristas de G. Además, la función de incidencia es la misma (Def. 4.11).

Definición 4.11 (Subgrafo) *Dado un grafo $G = (V, E, \varphi)$, $G' = (V', E', \varphi')$ es subgrafo de G sii $V' \subseteq V \wedge E' \subseteq E \wedge \varphi' \subseteq \varphi$.*

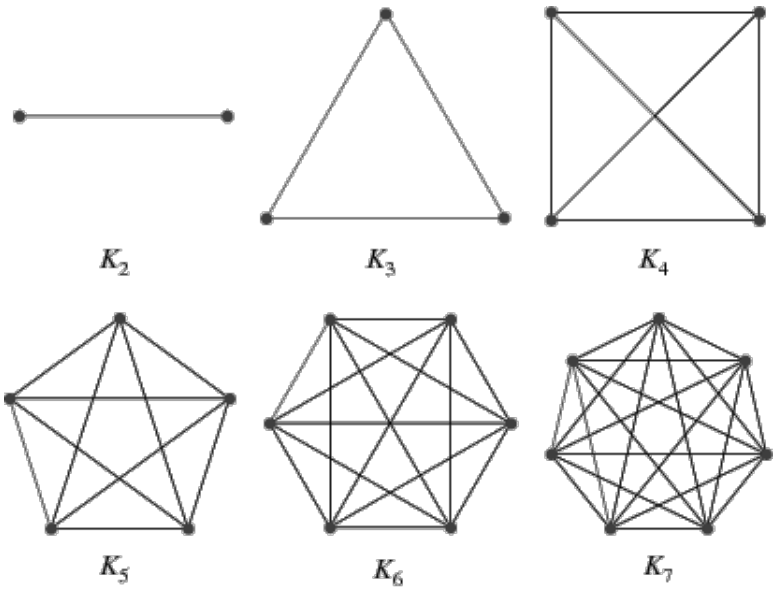


Figura 4.2: Ejemplo de grafos completos

Por último, un clique se define como se detalla en la definición 4.12.

Definición 4.12 (Clique) *Se dice que C es un clique de un grafo G sii C es un subgrafo de G y C es completo.*

Además, se dice que C es un clique maximal si no existe un clique de mayor tamaño en G, esto es con un mayor número de vértices, que lo contenga.

Por ejemplo, la subfigura 4.3b remarca mediante el color rojo y lineas punteadas un subgrafo completo K_3 del grafo de la figura 4.3a, constituyendo un clique de tamaño 3. Sin embargo, este clique no es máximo puesto que forma parte de un clique de tamaño 4, el cual se puede observar en la subfigura 4.3c.

4.1.1.3. Digrafos

Por otro lado, otra estructura de interés son los digrafos, que permiten asignar un "sentido" a las aristas, que pasan a denominarse "arcos" en este contexto. Cada arco posee exactamente un vértice de origen y exactamente un vértice de destino.

Formalmente, entonces, un difrafo se define como sigue:

Definición 4.13 (Digrafo) *Un digrafo D es una tupla (V, A, φ), donde V es un conjunto finito no vacío de vértices (llamados también nodos), A es un conjunto finito de*

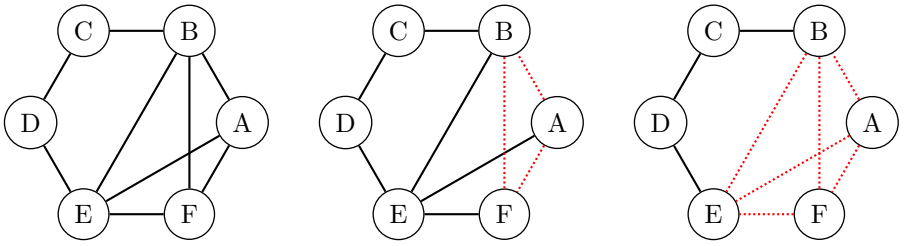


Figura 4.3: Ejemplo que ilustra la diferencia entre un clique y un clique máximo. (a) Grafo de ejemplo; (b) Se remarca en rojo y líneas punteadas un clique con 3 elementos que no es máximo pues es subgrafo de un clique de órden superior; (c) se remarca en rojo un clique máximo, pues no es subgrafo de un clique de orden superior.

arcos, y $\varphi : A \to V \times V$ *es la función de incidencia que especifica qué par ordenado de vértices le corresponde a cada arco.*

Como se puede apreciar en 4.1b, el conjunto de vértices de incidencia ahora posee un orden dado, siendo el primer vértice del par ordenado el vértice de "origen" y el segundo el vértice de "destino" del arco.

Como consecuencia de esta variación en la definición, la terminología asociada a los digrafos cambia de la siguiente manera:

Definición 4.14 (Incidencia positiva y negativa) *Dados* $a \in A$ *y* $v \in V$*, se dice que:*

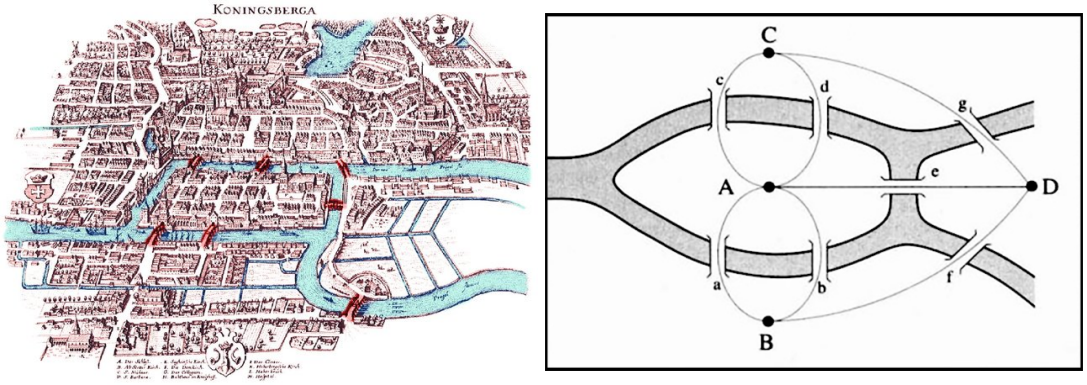
- a *incide positivamente sobre* $v \Leftrightarrow \varphi(a) = (o, d) \wedge o = v$
- a *incide negativamente sobre* $v \Leftrightarrow \varphi(a) = (o, d) \wedge d = v$

Definición 4.15 (Adyacencia en digrafos) *Dado un digrafo* $D = (V, A, \varphi)$ *y dos vértices* $v_1 \in V$ *y* $v_2 \in V$*, se dice que*

- v_1 *es adyacente a* $v_2 \Leftrightarrow \exists a \in A, \varphi(a) = (v_1, v_2)$
- v_2 *es adyacente desde* $v_1 \Leftrightarrow \exists a \in A, \varphi(a) = (v_1, v_2)$

Definición 4.16 (Arcos estrictamente paralelos) *Dados dos arcos cualesquiera* a_1 *y* a_2 *de un digrafo, se dice que* a_1 *es estrictamente paralelo a* $a_2 \Leftrightarrow \varphi(a_1) = \varphi(a_2)$

Definición 4.17 (Bucles en digrafos) *Dado un arco cualquiera* a *de un digrafo, se dice que* a *es un bucle* $\Leftrightarrow \varphi(a) = (v_1, v_2) \wedge v_1 = v_2$



(a) (b)

Figura 4.4: (a) Imagen ilustrativa de la ciudad de Kaliningrado, Rusia, en la que se situa el problema de los siete puentes. Imagen extraída de Acatitla Romero and Urbina Alonso (2017). (b) Modelo del problema mediante grafos. Imagen extraída de Gilkey (2019).

4.1.1.4. Grafos y digrafos como modelos

Los grafos y los digrafos son una poderosa herramienta que nos permiten representar situaciones complejas con una estructura que puede analizarse tanto visualmente y con la ayuda de una computadora. Como con todos los objetos matemáticos, es posible utilizar grafos y digrafos para realizar abstracciones: representación de estructuras de la realidad, representación de conocimiento, situaciones de flujo, entre otros (Rosen and Krithivasan, 2012).

Un famoso ejemplo de modelo mediante grafos es el caso del problema de los puentes de Königsberg o problema de los siete puentes de Königsberg, resuelto por L. Euler en 1736, que da origen a toda esta teoría. El mismo se situa en la ciudad de Königsberg (actual Kaliningrado, Rusia), la cual es atravesada por el río Pregel que se bifurca creando una isla entre sus brazos. Las regiones separadas por el río se unen mediante siete puentes, tal como se ilustra en la figura 4.4a. El problema consiste en hallar un recorrido para caminar por todas las regiones de la ciudad pasando por cada puente solo una vez y finalizando en la misma región de origen. Mediante la construcción de un grafo que represente esta situación (4.4b) pueden analizarse sus propiedades para determinar si existe o no este circuito. Euler resuelve este problema en Euler (1741) al determinar que solo existe este camino si todos los vértices del grafo poseen grado par.

Los grafos también pueden ser utilizados para modelar estructuras sociales más

complejas basadas en distintas relaciones que se dan entre personas o grupos de personas. En estos modelos los vértices representan individuos u organizaciones mientras que las aristas representan las relaciones entre los mismos. Generalmente tanto los vértices como las aristas se encuentran etiquetados con propiedades de los objetos modelados.

Asimismo, se ha observado en estudios de comportamiento de grupos que ciertas personas pueden influir en la forma en la que piensan otras. Puede usarse un grafo dirigido, llamado grafo de influencias, para representar este comportamiento. En la actualidad suele escucharse el término "*influencer*" en el contexto de las redes sociales, haciendo referencia a este fenómeno.

Estas y otras regularidades pueden ser extraídas de grafos y digrafos haciendo uso de la minería de grafos, cuestión que es abordada en la siguiente sección.

4.1.1.5. Minería de grafos

Debido a que los grafos son estructuras convenientes para modelar situaciones del mundo real de una manera fiel y cercana al mismo, la obtención de conocimiento que involucre interacciones entre distintos elementos de un dominio puede realizarse a partir de estos modelos.

Las regularidades que pueden observarse en grafos son plurales y sus aplicaciones se dan en diversos campos como la criminología, por ejemplo para la detección de asociaciones con hechos delictivos, la medicina, para la búsqueda de patrones de propagación de virus, planeamiento urbano, para la exploración de comportamiento de residentes de una ciudad en relación al transporte público, entre varias otras (Barabási, 2003; Barabási et al., 2016; Chen et al., 2003; Pastor-Satorras and Vespignani, 2001; Wang et al., 2017).

En general se pueden dar dos situaciones que dependen de lo modelado: O bien la base de datos está conformada por múltiples pequeños grafos, o bien la misma se compone por un único grafo de grandes dimensiones. El primer caso es habitual en aplicaciones como la biología o la química (e.g. enlaces químicos, relaciones entre especies, etc.), mientras que el segundo caso se nota generalmente en dominios donde destaca la cantidad de relaciones entre los vértices, como se da en redes sociales, por ejemplo.

Este tipo de estructuras matemáticas son difíciles de explorar cuando la cantidad de

nodos y aristas se magnifica. Consecuentemente, la minería de grafos como disciplina surgió oportunamente para solucionar estos problemas mediante el uso de técnicas informáticas para revelar regularidades de antigua definición que resultan de interés para la toma de decisiones (Washio and Motoda, 2003).

Entre estas regularidades se pueden enumerar algunas de las más abordadas en la literatura: la búsqueda de isomorfismos en pequeños grafos, la determinación del máximo subgrafo común, la búsqueda de subgrafos frecuentes, los subconjuntos de items frecuentes, las secuencias o caminos frecuentes, la búsqueda de cliques máximos, etc. (Aggarwal et al., 2010).

En esta tesis doctoral haremos uso de algoritmos para la búsqueda de regularidades al modelar las bases de datos espaciales y las relaciones entre estos datos mediante grafos, tal como se describe más adelante en las secciones correspondientes.

4.1.2. Teoría de conjuntos difusos

En esta sección se recuerdan brevemente los conceptos asociados a la teoría de conjuntos clásica o conjuntos rígidos (Sección 4.1.2.1) y se introduce la teoría de conjuntos difusos contrastando con la anterior (Sección 4.1.2.2).

4.1.2.1. Introducción

La teoría de conjuntos difusos nació en el año 1965 a partir del trabajo del profesor Lofti A. Zadeh, con el propósito de describir y formalizar la realidad mediante modelos flexibles que incluyan datos inciertos o vagos para los que la dicotomía “pertenece - no pertenece” resulta demasiado estricta. Es así como mediante la incorporación de un grado de pertenencia a conjuntos permite representar razonamientos con predicados imprecisos.

La teoría de conjuntos clásica o de conjuntos “nítidos” establece la pertenencia a un conjunto (representada con el símbolo $\in$) como un concepto primitivo que relaciona un elemento de un universal dado con un conjunto definido en dicho universal.

Los conjuntos se definen a través de los elementos que pertenecen al mismo, siendo dos conjuntos iguales si y solo si poseen exactamente los mismos elementos, sin importar orden o cantidad de apariciones de cada elemento. Esto se conoce como Axioma de Extensión (Def. 4.18).

Definición 4.18 (Axioma de extensión) *Dado un universal* $\mathcal{U}$, $\forall A, B \subseteq \mathcal{U} : \forall x, (x \in A \iff x \in B) \Rightarrow A = B$

Si se cumple que, dados dos conjuntos A y B, todo elemento que pertenece a A pertenece también a B entonces decimos que A se encuentra incluido en B ($A \subseteq B$) o que A es un subconjunto de B (Def. 4.19). El conjunto de todos los subconjuntos posibles de un conjunto dado se denomina conjunto potencia o potencial ($\wp(A)$) (Def. 4.20).

Definición 4.19 (Subconjunto) *Dado un universal* $\mathcal{U}$ *y dos conjuntos* $A \subseteq \mathcal{U}$ *y* $B \subseteq \mathcal{U}$, $A \subset B \iff \forall x \in A : x \in B$

Definición 4.20 (Potencial de un conjunto.) *Dado un universal* $\mathcal{U}$ *y un conjunto* $A \subseteq \mathcal{U}$, $\wp(A) = \{T : T \subseteq A\}$.

Debido a la naturaleza binaria de la pertenencia de un elemento a un conjunto, es posible utilizar la lógica bivaluada para representar de una manera cómoda y general esta relación. Esta representación se lleva a cabo mediante el uso de la denominada función característica del conjunto, la cual asigna a cada elemento del universal de referencia un valor 1 o 0 si el mismo pertenece o no al conjunto en cuestión (Def. 4.21).

Definición 4.21 (Función característica) *Dado un universal* $\mathcal{U}$ *y un conjunto* $A \subseteq \mathcal{U}$,

$$\mu_A : \mathcal{U} \rightarrow \{0,1\} / \begin{cases} \mu_A(x) = 1 \textit{ si } x \in A \\ \mu_A(x) = 0 \textit{ si } x \notin A \end{cases}$$

Las distintas formas de operar sobre los conjuntos pueden ser descriptas en términos de pertenencia o en términos de operaciones matemáticas sobre los valores de pertenencia de los elementos del universal. A modo de ejemplo, se definen a continuación las operaciones básicas de la teoría de conjuntos nítidos mediante estas dos modalidades equivalentes (Def. 4.22, Def. 4.23).

Definición 4.22 (Operaciones entre conjuntos nítidos) *Dados dos conjuntos* $A \subseteq \mathcal{U}$ *y* $B \subseteq \mathcal{U}$*:*

- *Unión:* $A \cup B = \{x / x \in A \vee x \in B\}$
- *Intersección:* $A \cap B = \{x / x \in A \wedge x \in B\}$

- *Complemento de A:* $\bar{A} = \{x/x \notin A\}$

Definición 4.23 (Operaciones mediante valores de pertenencia) *Dados dos conjuntos* $A \subseteq \mathcal{U}$ *y* $B \subseteq \mathcal{U}$*:*

- *Unión:* $\mu_{A \cup B}(x) = max\{\mu_A(x), \mu_B(x)\}$
- *Intersección:* $\mu_{A \cap B}(x) = min\{\mu_A(x), \mu_B(x)\}$
- *Complemento de A:* $\mu_{\bar{A}}(x) = 1 - \mu_A(x)$

4.1.2.2. Conjuntos difusos

Supongamos ahora un universal conformado por un conjunto de personas del cual queremos distinguir un subconjunto de personas "altas", siendo el predicado "x es alto" un predicado impreciso o vago cuyo valor de verdad dependerá no sólo de la persona, sino también del contexto considerado: si tomamos como altas a las personas cuya altura es 1,80m, ¿resulta realista excluir a alguien que mide 1,79m? Estos predicados son habituales en la comunicación de conceptos cotidianos por lo que resultan altamente informativos e importantes para la toma de decisiones.

La pertenencia a un conjunto dado definido por un predicado impreciso como el mencionado en el párrafo anterior, entonces, es cuestión de grados entre los extremos "pertenece" y "no pertenece". De esta forma la función característica de un conjunto puede tomar cualquier valor en un intervalo continuo $[0, 1]$ (Def. 4.24). Retomando el ejemplo de la sección anterior, una persona puede ser alta con un grado de pertenencia al conjunto igual a 0.8 si "es alta pero no tanto".

Definición 4.24 (Conjunto difuso) *Dado un universal* $\mathcal{U}$ *continuo o discreto, un subconjunto difuso* $\tilde{A} \subset \mathcal{U}$ *es una función* $\mu_{\tilde{A}} : \mathcal{U} \rightarrow [0, 1]$ *que asigna a cada elemento del universal un grado de pertenencia de un elemento x a* $\tilde{A}$*.*

Consecuentemente, resulta necesario redefinir los conceptos de igualdad e inclusión que hemos mencionado en la sección anterior. Para esto haremos uso exclusivo de los valores o grados de pertenencia obtenidos a partir de la función característica de los conjuntos difusos de la forma que se detalla a continuación.

Definición 4.25 (Igualdad de conjuntos difusos) *Dados dos conjuntos difusos* $\tilde{A} \subset \mathcal{U}$ *y* $\tilde{B} \subset \mathcal{U}$*,* $\tilde{A} = \tilde{B} \iff \forall x \in \mathcal{U} : \mu_{\tilde{A}}(x) = \mu_{\tilde{B}}(x)$

Definición 4.26 (Inclusión de conjuntos difusos) *Dados dos conjuntos difusos* $\tilde{A} \subset \mathcal{U}$ *y* $\tilde{B} \subset \mathcal{U}$, $\tilde{A} \subset \tilde{B} \iff \forall x \in \mathcal{U} : \mu_{\tilde{A}}(x) \leq \mu_{\tilde{B}}(x)$

No obstante, las operaciones definidas sobre los conjuntos nítidos mediante el uso de grados de pertenencia pueden aplicarse de forma idéntica sobre conjuntos difusos.

Resulta necesario además definir dos conceptos adicionales: el soporte de un conjunto difuso y el subconjunto nítido de nivel α o α-corte. El soporte de un conjunto difuso es el conjunto nítido formado por todos los elementos que pertenecen al conjunto difuso en algún grado no nulo (Def. 4.27). Un α-corte, por otro lado, es el conjunto nítido formado por todos los elementos que pertenecen al conjunto difuso con un grado de pertenencia mayor a un valor α dado (Def. 4.28). Resulta evidente, pues, que un corte de nivel 0 es igual al soporte de la relación difusa.

Definición 4.27 (Soporte de un conjunto difuso) *Dado un universal* $\mathcal{U}$ *y un subconjunto difuso* $\tilde{A} \subset \mathcal{U}$, *el soporte de un conjunto difuso* $(\tilde{A}) = \{x/x \in \mathcal{U} \wedge \mu_{\tilde{A}}(x) > 0\}$

Definición 4.28 (Corte de nivel α o α-corte) *Dado un universal* $\mathcal{U}$ *y un subconjunto difuso* $\tilde{A} \subset \mathcal{U}$, *el* α*-corte de conjunto difuso* $A_\alpha = \{x/x \in \mathcal{U} \wedge \mu_{\tilde{A}}(x) > \alpha\}$ *para* $\alpha \in [0, 1)$

Por último, se han establecido modelos genéricos para las operaciones binarias de intersección y unión sobre la teoría de conjuntos difusos (Menger, 1942). Estas operaciones se denominan T-norma o Norma Triangular (Def. 4.29) y T-conorma o Conorma Triangular (Def. 4.30) respectivamente y cumplen con las propiedades de conmutatividad, asociatividad, monotonicidad y condiciones de frontera.

Definición 4.29 (T-norma) *Operación binaria* $\top : [0, 1]^2 \to [0, 1]$ *tal que:*

- $x \top y = y \top x$ *(Conmutatividad)*
- $x \top (y \top z) = (x \top y) \top z$ *(Asociatividad)*
- $x \leq y \wedge w \leq z \to x \top w \leq y \top z$ *(Monotonicidad)*
- $x \top 0 = 0$ *(Elemento absorbente)*
- $x \top 1 = x$ *(Elemento neutro)*

Definición 4.30 (T-conorma) *Operación binaria* $\perp : [0, 1]^2 \to [0, 1]$ *tal que:*

- $x \perp y = y \perp x$ *(Conmutatividad)*
- $x \perp (y \perp z) = (x \perp y) \perp z$ *(Asociatividad)*
- $x \leq y \wedge w \leq z \rightarrow x \perp w \leq y \perp z$ *(Monotonicidad)*
- $x \perp 0 = x$ *(Elemento neutro)*
- $x \perp 1 = 1$ *(Elemento absorbente)*

Casos particulares de T-norma y T-conorma son la ya presentadas T-norma mínimo y T-conorma máximo, enunciadas para conjuntos rígidos pero aplicables a conjuntos difusos (Def. 4.23). Ejemplo de otras t-normas como son la T-norma Producto, la T-norma Producto Drástico, la T-norma Producto de Einstein y la T-norma Producto de Hamacher pueden observarse en las ecuaciones 4.1, 4.2, 4.3, 4.4 y 4.5. Asimismo, una comparativa del valor de estas funciones puede verse en la figura 4.5.

$$\top_{min}(a, b) = Min(a, b) \tag{4.1}$$

$$\top_{prod}(a, b) = a * b \tag{4.2}$$

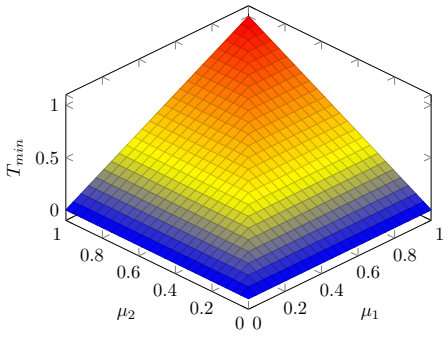
$$\top_{drast}(a, b) = \begin{cases} x & \text{si } y = 1 \\ y & \text{si } x = 1 \\ 0 & \text{en caso contrario} \end{cases} \tag{4.3}$$

$$\top_{E}(a, b) = (a \cdot b) \cdot (1 + (1 - x) + (1 - y))^{-1} \tag{4.4}$$

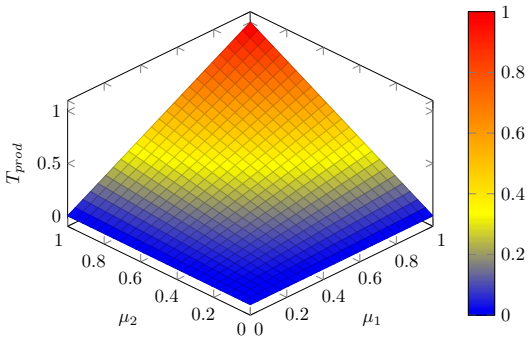
$$\top_{H_\lambda}(a, b) = \begin{cases} 0 & \text{si } \lambda = a = b = 0 \\ (a \cdot b) \cdot (\lambda + (1 - \lambda) \cdot (a + b - a \cdot b))^{-1} & \text{en caso contrario} \end{cases} \tag{4.5}$$

4.2. Procesos de explotación de información espacial

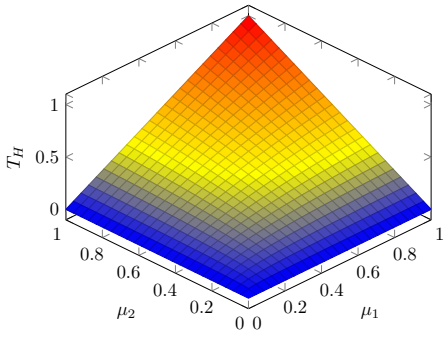
En la presente sección se describen cuatro procesos de explotación de información espacialmente referenciada. Estos procesos se relacionan con actividades de la minería



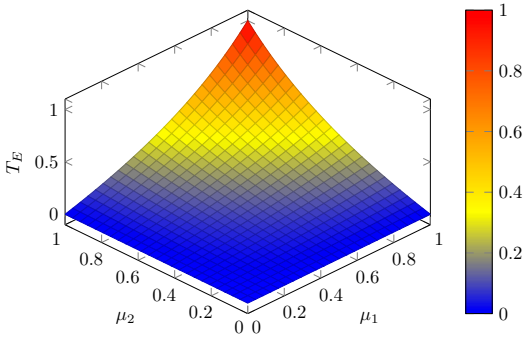
(a) T-norma mínimo



(b) T-norma producto



(c) T-norma producto de Hamacher



(d) T-norma producto de Einstein

Figura 4.5: Comparación de las T-normas mínimo y producto de Hamacher.

de datos espaciales y combinan algoritmos tanto de esta disciplina como de la minería de datos tradicional y minería de grafos para obtener conocimiento que resulte útil para la toma de decisiones estratégica en distintos escenarios.

Los procesos están pensados para datos espacialmente referenciados en formato vectorial, por lo que en caso de disponer de bases de datos *raster* resulta necesaria una traducción previa, tal como se detalla en la sección 2.4.3. Cualquiera de los tres tipos básicos de datos espaciales vectoriales son factibles de ser usados como entrada, aunque algunos procesos tienden a ser utilizados preferentemente con solo un subconjunto de los mismos debido a las características de las situaciones modeladas.

Entre los cuatro procesos en cuestión se encuentra un proceso para el descubrimiento de grupos espaciales, que generaliza y sistematiza la aplicación de algoritmos de *clustering* sobre este tipo particular de datos; un proceso para el descubrimiento de anomalías espaciales, que considera la detección de anomalías en contextos locales previamente definidos o a definir para reducir la heterogeneidad espacial y permite la determinación de zonas de comportamiento anómalo grupal; un proceso para el descubrimiento de asociaciones espaciales, que permite, al igual que el proceso anterior,

disponer de patrones locales y modelar múltiples relaciones espaciales simultáneas ya sean rígidas o difusas; y por último un proceso para el descubrimiento de patrones de co-localización, que puede verse como un caso particular del proceso de descubrimiento de asociaciones espaciales.

Estos procesos esquematizan las actividades a ser realizadas sin especificar exactamente cómo han de ser implementadas, pudiendo optar por distintas estrategias, algoritmos o incluso por el conocimiento de expertos para conseguir los resultados esperados. Adicionalmente, algunas secciones de estos algoritmos pueden ser reemplazadas o removidas en caso de que no se adapten a la situación particular que se esté abordando.

A continuación se detallan cada uno de los mismos en el orden en el que se los introdujo previamente.

4.2.1. Proceso de descubrimiento de grupos espaciales

El proceso de agrupamiento de datos espaciales aplica cuando se desean obtener subconjuntos de la masa de información espacialmente referenciada, sea esta formada por puntos o regiones vectoriales, y determinar las características en común de estos subgrupos previamente desconocidos.

Son ejemplos de aplicación determinar características demográficas comunes en secciones administrativas, hallar las características de altas concentraciones de eventos en el espacio tales como focos de incendios forestales, segmentar elementos geográficos tales como ciudades o puntos de entrega con características comunes con propósitos de logística, entre otros.

Este proceso consta de tres subprocesos, tal como se puede observar en la figura 4.6: preparación de los datos, agrupamiento de datos espaciales y descripción de los grupos.

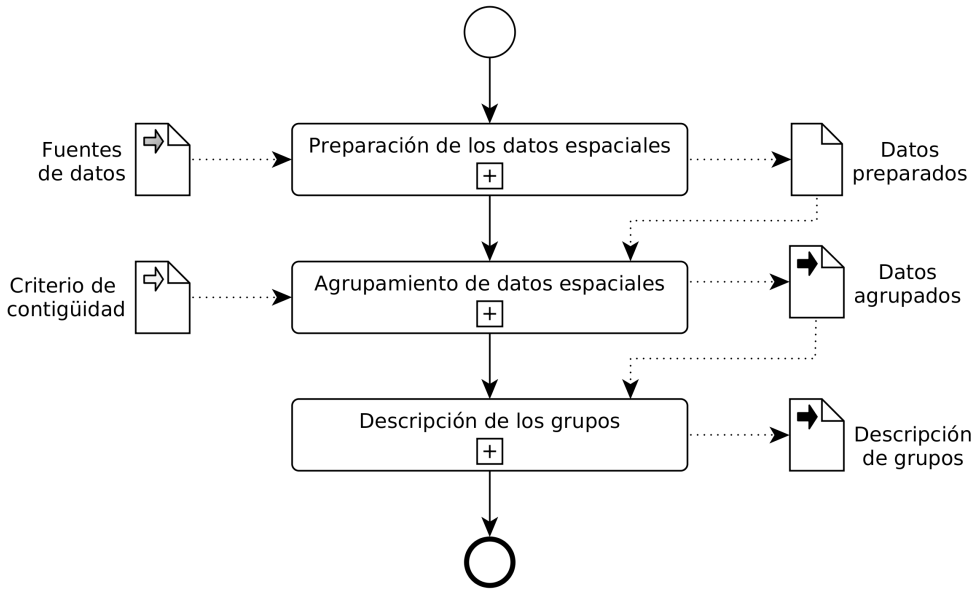


Figura 4.6: Proceso de descubrimiento de grupos espaciales

4.2.1.1. Preparación de los datos

La preparación de los datos es una actividad contemplada en todos los modelos de proyectos de explotación de información, tal como se ha mencionado en la sección 2.2. Esta toma como entrada la o las fuentes de datos crudas y devuelve un conjunto de datos de calidad para la obtención de patrones de conocimiento.

Esta actividad depende no solo de la calidad de los datos, la cual debe ser garantizada antes de operar sobre los registros, sino también del problema de negocio que se está abordando y, por consiguiente, de los tipos de algoritmos de minería de datos que se utilizarán de forma posterior: algoritmos de agrupamiento para detección de

Algoritmo 4.1 Proceso de descubrimiento de grupos espaciales

Entrada: $Datos \leftarrow$ Fuente de datos espaciales
Entrada: [OPT] $Crit \leftarrow$ Criterio de contigüidad
Salida: Datos espaciales agrupados
Salida: Descripción de los grupos
1: $D_p \leftarrow$ Preparar datos espaciales($Datos$)
2: $G \leftarrow$ Agrupar los datos espaciales(D_p, [OPT] $Crit$)
3: $Desc \leftarrow$ Describir los grupos(G)
4: **devolver** G
5: **devolver** $Desc$

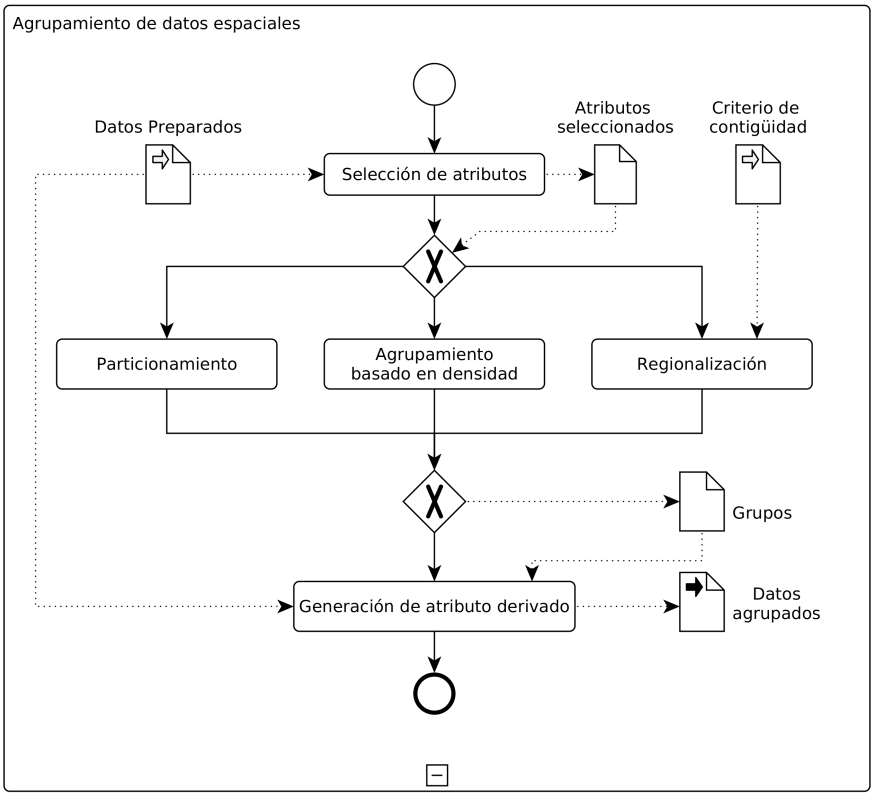


Figura 4.7: Subproceso de agrupamiento de datos espaciales

grupos espaciales, para detección de zonas calientes o para detección de regiones. Si bien mencionamos este detalle en este apartado haremos hincapié en las características de los atributos de los datos según el tipo de algoritmo de problema de requisito en el apartado siguiente.

No obstante, una consideración sobre los datos de entrada debe de ser tenida en cuenta: no resulta factible segmentar objetos vectoriales de distinto tipo. La aplicación de este proceso se realiza generalmente sobre un tipo uniforme de datos espaciales cuya selección depende de las características a modelar del dominio del problema. Por ejemplo, en una base de datos de divisiones administrativas geopolíticas, o se busca agrupar ciudades (puntos), o regiones (polígonos), u otras divisiones administrativas. Por otro lado, desde un punto de vista semántico los resultados de agrupar objetos espaciales disimiles no sería congruente para un problema particular aún cuando el objeto vectorial seleccionado para representarlos sea el mismo.

4.2.1.2. Agrupamiento de datos espaciales

El subproceso de agrupamiento de datos espaciales (Fig. 4.7, Alg. 4.2) se encarga de la aplicación de algoritmos de segmentación para la obtención de subgrupos de datos con características comunes sobre un conjunto de datos espaciales provistos como entrada que hayan sido previamente preparados.

Primeramente, los atributos son seleccionados según el tipo de agrupamiento espacial que es requerido para dar respuesta a la inteligencia de negocio. En caso de agrupamiento para la detección de patrones de puntos o puntos calientes (*hotspots*) resulta relevante utilizar la posición de los objetos de la base de datos (puntos, en este caso) como entrada para el algoritmo, por lo que los atributos espaciales de los datos deberán ser seleccionados. En caso de que la autocorrelación espacial deba ser especialmente tenida en cuenta, es posible considerar utilizar algoritmos de agrupamiento basados en índices estadísticos como la I de Moran, por ejemplo (Getis and Ord, 2010). En este caso será necesario incluir no solo la posición y extensión de los objetos espaciales, sino también los valores de sus atributos no espaciales.

Por otro lado, si resultase necesaria la obtención de regiones espacialmente contiguas, es menester seleccionar tanto los atributos espaciales para asegurar la contigüidad de los datos (puntos o polígonos, generalmente) como así también los atributos no espaciales de los mismos para el cálculo de la similitud entre ellos. Cabe además mencionar que, debido a las diversas formas de determinar la contigüidad entre dos objetos espaciales, se requiere aportar como entrada del proceso el criterio de contigüidad adecuado según el problema de negocio que se está abordando.

Por último, en el caso de agrupamiento espacial, los atributos espaciales y los no espaciales no poseen distinción, por lo que cualquier subconjunto de ellos puede ser seleccionado y considerado para evaluar la similitud entre los elementos que conformarán los grupos.

Posteriormente, los datos son sometidos a un algoritmo de agrupamiento o *clustering* (Han et al., 2011) , el cual depende de igual forma que la selección de atributos del paso anterior. En caso de descubrimiento de patrones de puntos el proceso indica el uso de algoritmos de agrupamiento basados en densidad tales como aquellos de la familia DBSCAN (Ester et al., 1996). La selección de uno u otro algoritmo de esta familia dependerá de las distintas particularidades del algoritmo, de los datos o de disponibilidad de recursos técnicos que se posean para la implementación del mismo

(Dharni and Bnasal, 2013; Esfandani and Abolhassani, 2010; He et al., 2014; Miyahara and Miyamoto, 2014; Song and Lee, 2018; Vijayalaksmi and Punithavalli, 2012). La regionalización, por otro lado, se ejecuta haciendo uso de algoritmos de regionalización, los cuales tienen en cuenta la contigüidad entre puntos o polígonos espaciales. Se propone el uso de la familia de algoritmos REDCAP (Guo, 2008) para esta tarea.

Por último, en el caso de agrupamiento de datos, cualquier algoritmo de agrupamiento clásico como k-means o SOM (Kohonen, 1982) puede ser utilizado al considerar los datos espaciales y no espaciales de forma indistinta.

Debido a que estos algoritmos generalmente proveen resultados en distintos formatos, se debe integrar los grupos obtenidos al conjunto de datos original mediante un etiquetado en una nueva columna construida para tal finalidad. Como resultado de la aplicación de este subproceso se obtiene un conjunto de datos etiquetados con el grupo al cual pertenece cada uno de ellos.

Una vez hechos con la segmentación de los grupos, se procede con el proceso de descripción descripto a continuación.

Algoritmo 4.2 Subproceso de agrupamiento de datos espaciales

Entrada: $D_p \leftarrow$ Datos preparados
Entrada: [OPT] $Crit \leftarrow$ Criterio de contigüidad
Salida: Datos agrupados
1: $D_s \leftarrow$ Seleccionar atributos(D_p)
2: **si** Detección de patrones de puntos **entonces**
3: $Gs \leftarrow$Aplicar algoritmo DBSCAN(D_s)
4: **si no**, **si** Regionalización **entonces**
5: $Gs \leftarrow$ Aplicar algoritmo REDCAP(D_s, $Crit$)
6: **si no**
7: $Gs \leftarrow$ Aplicar algoritmo de agrupamiento(D_s)
8: **fin si**
9: $G \leftarrow$ Integrar grupos(D_p, Gs)
10: **devolver** G

4.2.1.3. Descripción de los grupos

La descripción de los grupos toma como entrada el conjunto de datos etiquetado en categorías correspondiente a los grupos obtenidos mediante la aplicación del subproceso de agrupamiento de datos espaciales (Fig. 4.8, Alg. 4.3).

Esta descripción tiene como objetivo explicitar los patrones internos de cada grupo

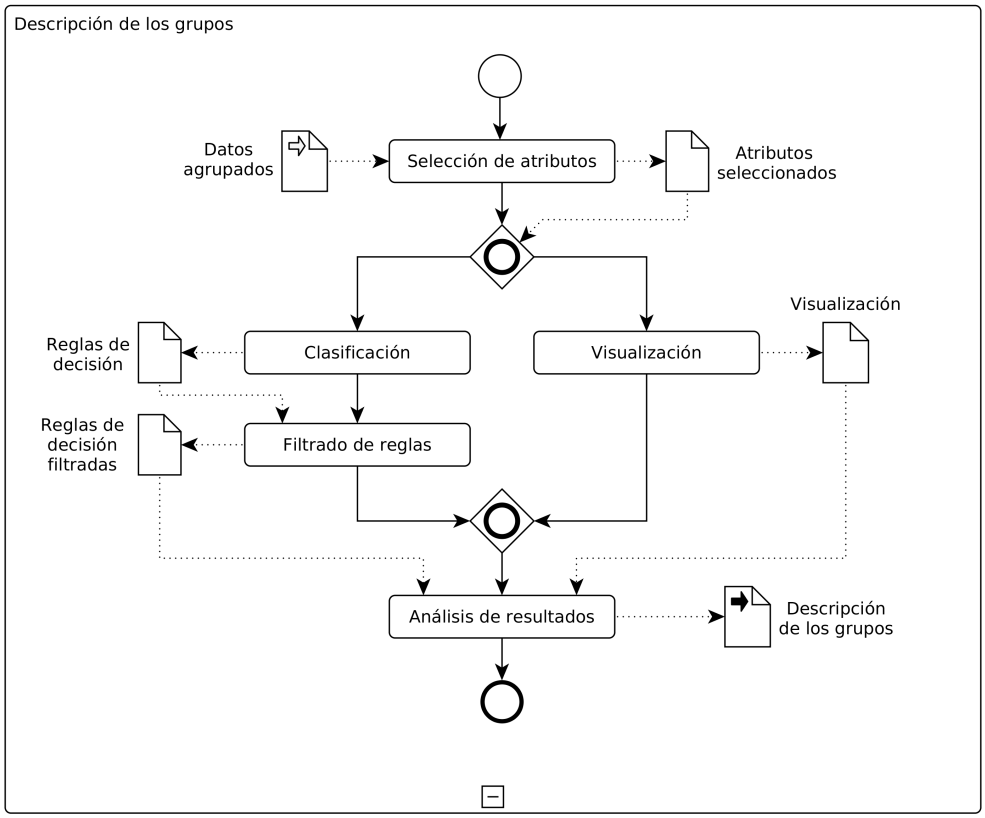


Figura 4.8: Subproceso de descripción de grupos espaciales

descubierto para brindar a los tomadores de decisiones información que sea de utilidad para esta tarea debido a que los datos categorizados, *a priori*, no aportan ningún valor para quienes se encuentran en funciones estratégicas dentro de la organización. Para esto, la coherencia interna de cada grupo debe ser elevada (Aggarwal, 2015).

Según el objetivo del agrupamiento, los algoritmos y los atributos considerados, distintos aspectos de la base de datos resultan interesantes de ser extraídos, por lo que el subproceso de descripción de los datos propone la utilización de tanto herramientas de visualización como algoritmos de clasificación para este propósito.

El primer paso del subproceso consiste en determinar los atributos involucrados. La clasificación de datos generalmente se realiza sobre los atributos no espaciales de los elementos de la base de datos, los cuales son los que intervienen en la evaluación de la similitud en el subproceso de agrupamiento, específicamente al utilizar agrupamiento o regionalización.

En el caso de detección de patrones de puntos, en el que no se ven involucrados los atributos no espaciales, el uso de los mismos tiene sentido si el propósito es determinar

las características en común que tienen los *hotspots* mas es altamente probable que no sea posible obtener una alta precisión en la clasificación obtenida. Esto se debe a que los grupos habrían sido determinados según un criterio de densidad y no por la similitud de los mismos. En este caso, por lo regular, se recurre a la visualización. En ambos casos, el atributo objetivo es el atributo generado en el que consta a qué grupo pertenece cada elemento del conjunto de datos

Una vez seleccionados los atributos dos caminos son posibles: la descripción de los grupos mediante reglas, tal como se realiza en el proceso de explotación de información para el descubrimiento de reglas de pertenencia a grupos (Britos, 2008), o la visualización de los datos clasificados. Estos dos caminos no son exclusivos, tal como se observa en el diagrama de la figura 4.8, pudiéndose recurrir a ambas estrategias.

El proceso propone la utilización de reglas de clasificación obtenidas mediante algoritmos de aprendizaje supervisado basados en árboles como los previamente mencionados ID3, C4.5, C5.0, entre otros. Sin embargo, es posible hacer uso de descriptores estadísticos de tendencia central o de dispersión para tal tarea. Las reglas de decisión además pueden ser filtradas para obtener aquellas de mayor confianza o de mayor interés para la inteligencia de negocio. En cualquier caso, se buscan modelos altamente descriptivos.

Asimismo, la visualización de datos se puede realizar considerando la dimensión espacial por un lado y la dimensión no espacial por otro. La dimensión espacial considera generalmente la distribución de los objetos en el espacio, por lo que decidir el tipo de gráfico que representa mejor esta distribución resulta clave: ejes de coordenadas, mapas, etc. Las dimensiones no espaciales y las categorías de los grupos, por otro lado, serán expresados haciendo uso de códigos visuales adecuados para cada tipo de dato: color, tamaño, forma, etc. (Grinstein and Ward, 2002).

Por último, resulta clave el análisis de los resultados para determinar su confianza y la relevancia de los mismos para la resolución del problema de negocio, generando un documento de descripción de los grupos como salida del subproceso, en el que consta toda la información generada en esta etapa.

4.2.2. Proceso de descubrimiento de anomalías espaciales

El proceso de descubrimiento de anomalías espaciales (Fig. 4.9) se puede utilizar para hallar elementos de la base de datos con comportamiento que desvía del comporta-

Algoritmo 4.3 Subproceso de descripción de grupos espaciales

Entrada: $D_a \leftarrow$ Datos agrupados
Salida: Descripción de los grupos
1: $A_e \leftarrow$ Seleccionar atributos de entrada
2: $A_t \leftarrow$ Seleccionar atributo objetivo(D_t)
3: **si** Descripción mediante reglas **entonces**
4: $R \leftarrow$ Algoritmo TDIDT(A_e,A_t)
5: $R_f \leftarrow$ Filtrar reglas (R)
6: **fin si**
7: **si** Descripción mediante visualización **entonces**
8: $V \leftarrow$ Generar visualización(A_e,A_t)
9: **fin si**
10: $Desc \leftarrow$ Análisis de resultados (R_f, V)
11: **devolver** $Desc$

miento normal y determinar las características del fenómeno que las genera. Ejemplos de escenarios de aplicación son: determinar zonas geográficas con comportamiento climatológico anormal, determinar regiones censales con características demográficas que difieren de la norma, encontrar zonas geológicas con concentraciones de minerales que difieren de lo esperado, entre otros. Este proceso permite además evaluar las anomalías en una región o vecindario del terreno y determinar si dichos vecindarios constituyen una anomalía en si mismos al comparar sus valores con los de sus vecinos.

El proceso toma como entrada un conjunto de datos vectoriales. El primer paso del mismo consiste en la preparación de la entrada para incrementar la calidad de los datos y dejarla en condiciones para el próximo paso: la definición de los vecindarios. Los datos vectoriales pueden ser ingresados en cualquiera de sus tres formatos básicos: puntos, líneas o polígonos, mas es habitual la determinación de anomalías puntuales o relacionadas a zonas bidimensionales del espacio considerado. Como salida de esta actividad se obtiene un conjunto de datos vectoriales preparados. El formato de salida es independiente al proceso, debido a que depende exclusivamente de las entradas requeridas por las herramientas utilizadas en la actividad que le sigue de forma inmediata.

Cómo dijimos en el párrafo anterior, la siguiente actividad permite la definición de vecindarios de datos. Estos vecindarios son regiones del espacio con coherencia interna en la distribución de los objetos espaciales dentro del mismo. Estas regiones pueden estar previamente definidas por el dominio del problema o, si no lo están, es posible definirlas mediante un subproceso de agrupamiento de datos espaciales (Fig.

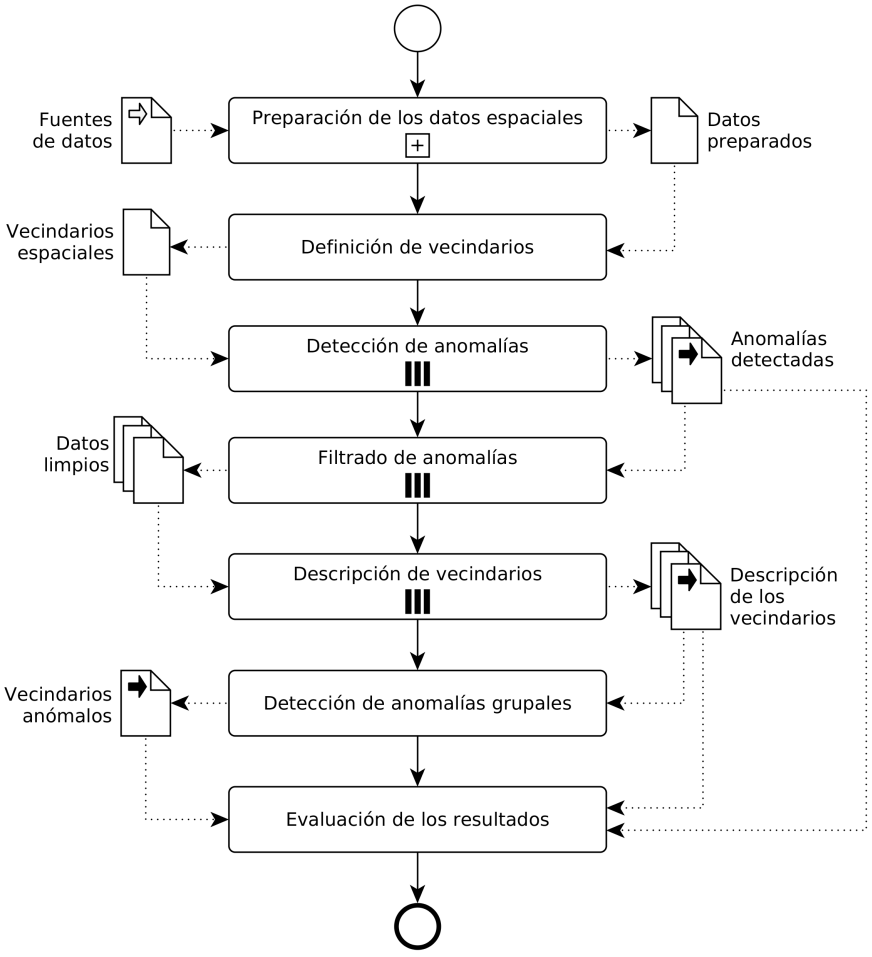


Figura 4.9: Proceso de descubrimiento de anomalías espaciales

4.2).

Se sugiere el uso de algoritmos basados en densidad para el agrupamiento para considerar la Primera Ley de la Geografía (sección 3.1) si se trata con datos vectoriales puntuales. En el caso de tratar con regiones se sugiere regionalizar utilizando un criterio de contigüidad basado en fronteras. Es posible además el uso de estrategias basadas en estadísticas para la determinación de zonas con datos autocorrelacionados espacialmente, si el problema lo requiere. Las líneas generalmente no son utilizadas en este proceso, mas no existe impedimento procedimental para hacerlo si se lo requiere.

Es menester aclarar que este paso es mandatorio aunque se defina un único vecindario, lo cual equivaldría a la búsqueda de patrones globales. Esto puede resultar de interés cuando la zona de búsqueda está acotada y su fragmentación no resulta interesante para la toma de decisiones. En este caso la actividad de detección de anomalías

Algoritmo 4.4 Proceso de descubrimiento de anomalías espaciales

Entrada: $Datos \leftarrow$ Fuente de datos espaciales
Salida: Anomalías detectadas
Salida: Descripción de los vecindarios
Salida: Vecindarios anómalos
1: $D_p \leftarrow$ Preparar datos espaciales($Datos$)
2: $G \leftarrow$ Definir vecindarios(D_p)
3: $A \leftarrow []$
4: $D \leftarrow []$
5: **para todo** $g \in G$ **hacer**
6: $A_g \leftarrow$ Detectar anomalías(g)
7: $A \leftarrow +A_g$
8: $L_g \leftarrow$ Filtrar datos anómalos(g,A_g)
9: $D_g \leftarrow$ Describir vecindario(L_g)
10: $D \leftarrow +D_g$
11: **fin para**
12: $VA \leftarrow$ Detectar anomalías grupales(Ds)
13: **devolver** A
14: **devolver** Ds
15: **devolver** VA

grupales explicada a continuación no tendría razón para ser ejecutada.

La definición de vecindarios, entonces, da como resultado un agrupamiento del conjunto de datos. A partir de esta segmentación se procede a realizar la detección de anomalías locales mediante la aplicación de algoritmo dedicados a tal fin, como LOF (Breunig et al., 2000) o SLOM (Chawla and Sun, 2006), entre otros (Hodge and Austin, 2004; Rottoli et al., 2018), sobre los atributos no espaciales de los datos.

Se propone considerar el uso de la distancia de Mahalanobis para calcular cuánto difieren los valores de los atributos no espaciales de la tendencia central del vecindario, debido a que este método considera correlaciones entre los valores de sus atributos (De Maesschalck et al., 2000). No obstante, la selección del método, como siempre, depende de las necesidades que derivan del problema abordado. Se obtiene de esta actividad un conjunto de entidades de datos en el formato que se requiera en fases posteriores.

Luego de la detección de los datos anómalos se debe limpiar la base de datos de estos valores apartándolos en otro soporte. El conjunto de datos limpio, entonces, será utilizado para describir cada uno de los vecindarios al no estar sesgados por los *outliers*. Para esto es plausible utilizar las técnicas propuestas en el subproceso de descripción

de grupos mencionado en el capítulo anterior (Fig. 4.8). No obstante, resulta más conveniente el uso de técnicas que permitan generar un conjunto de descriptores que puedan aportar información sobre el comportamiento general de cada vecindario de tal forma que en fases siguientes los mismos colaboren con la diferenciación de los grupos que desvían en gran medida de la normalidad de los mismos.

De esta forma, a partir de la tabla de descriptores generadas para cada región se aplica un método para el análisis de anomalías como los mencionados con anterioridad. De esta forma serán evidentes las regiones con comportamiento anómalo.

Por último, tanto las anomalías detectadas, la descripción de los vecindarios y los vecindarios anómalos son interpretados para dar respuesta a las necesidades del negocio planteadas. Las anomalías locales han de ser contrastadas con las descripciones de los grupos para determinar bajo que aspectos estas difieren de la normalidad de los datos. Distintas herramientas de visualización pueden dar soporte a esta tarea. De la misma forma, los valores anormales grupales deben compararse con los datos en sus vecindades espaciales.

Finalmente, distintas variantes pueden ser adoptadas en la implementación del proceso, como considerar una búsqueda de anomalías globales sin tener en cuenta la vecindad espacial. No obstante, estas consideraciones quedan supeditadas a las necesidades que el dominio del problema plantea y al juicio de los encargados de realizar la materialización del proceso.

4.2.3. Proceso de descubrimiento de asociaciones espaciales

El proceso de descubrimiento de asociaciones espaciales (Fig. 4.10) aplica cuando se desea encontrar subconjuntos de datos de interés para la inteligencia de negocios que se encuentran íntimamente relacionados espacialmente, implicando cierta relación causal posible entre los elementos de la base de datos (Róttoli et al., 2018).

Este proceso resulta útil para determinar regularidades en la distribución de edificios y servicios públicos en el área de planificación urbana, para determinar características complejas de patrones delictivos, para hallar posibles fuentes de contaminación ambiental, entre otros.

Se tiene en cuenta la localidad de los patrones de asociación y la posibilidad de modelar múltiples relaciones espaciales entre variados tipos de datos espaciales vectoriales mediante la utilización de la teoría de grafos para modelar estas interacciones

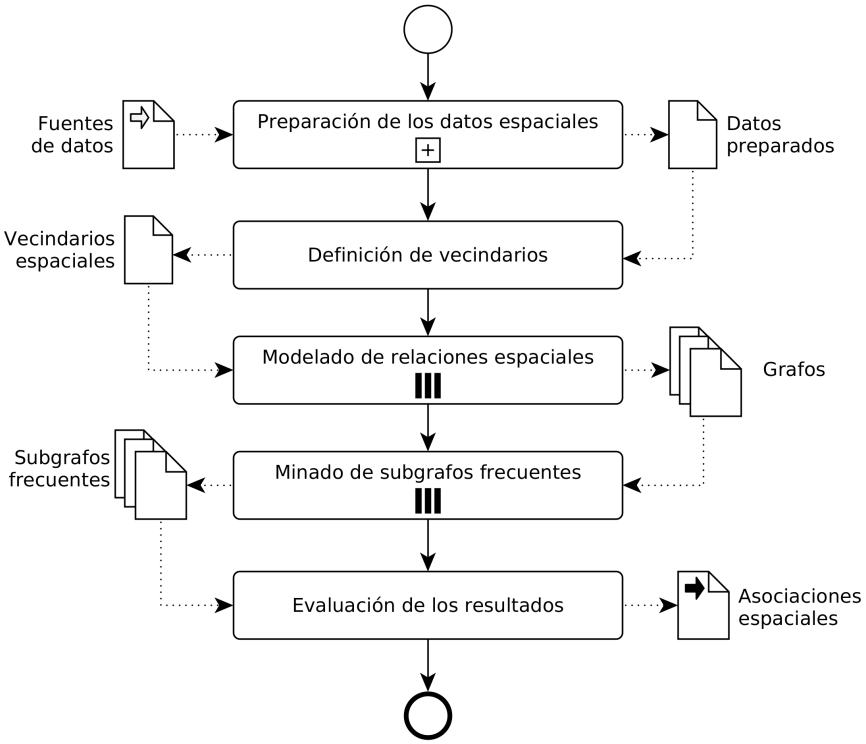


Figura 4.10: Proceso de descubrimiento de asociaciones espaciales

(sección 4.1.1).

El proceso comienza, como siempre, con la preparación de los datos espaciales. Estos deben contar con atributos espaciales y no espaciales diferenciados apropiadamente. En este proceso pueden coexistir múltiples tipos de datos distintos. Es factible además eliminar anomalías para evitar incorporar errores al proceso de descubrimiento.

Posteriormente resulta necesario definir los vecindarios en los cuales se buscarán las asociaciones entre los datos espaciales. Estos vecindarios pueden ser definidos por los tomadores de decisiones según información sobre el problema (por ejemplo, utilizando fronteras geopolíticas o separaciones por tipo de suelo, entre otros), o bien utilizar aprendizaje no supervisado mediante un subproceso de agrupamiento de datos espaciales (Fig. 4.2) para determinar concentraciones de puntos que puedan resultar de interés. En este último caso, al igual que en el proceso de descubrimiento de anomalías espaciales, se prefiere el uso de algoritmos basados en densidad debido a la Primera Ley de la Geografía (sección 3.1) o, de no ser posible, algoritmos de regionalización.

A partir de aquí, las siguientes actividades se realizan de forma paralela por cada

uno de los vecindarios obtenidos.

En primer lugar, se calculan y modelan las relaciones espaciales entre los distintos datos de interés para la inteligencia de negocio. Estas relaciones pueden ser de cualquier naturaleza: euclidianas, topológicas, direccionales o híbridas. Este paso puede ser costoso al definirse las relaciones binarias sobre un subconjunto de un producto cartesiano, tal como se describe en la sección 2.5.

Una vez calculadas, las relaciones espaciales pueden modelarse haciendo uso de grafos donde el conjunto de vértices está dado por las instancias de los objetos espaciales y la incidencia de las aristas se define mediante la existencia de un par ordenado en el conjunto formado por la unión de todas las relaciones espaciales calculadas, siendo además, un grafo etiquetado con los tipos de datos espaciales en los vértices y los tipos de relaciones espaciales en las aristas.

De esta forma, el grafo resultante está dado por la tupla $G = (V, E, L, K, \varphi, P_l, P_k)$, donde V es el conjunto de vértices que representan las instancias de los objetos espaciales, E es el conjunto de aristas que representan las relaciones entre las instancias de los objetos espaciales, L y K son los conjuntos de etiquetas para los vértices y las aristas respectivamente, representando los tipos de objetos espaciales y los tipos de relaciones, $\varphi : E \to \{\{v_1, v_2\} | v_1 \in V \wedge v_2 \in V \wedge v_1 \neq v_2\}$ es la función de incidencia de las aristas sobre los vértices, $P_l \subseteq V \times L$ es la relación de etiquetado de los objetos espaciales y $P_k \subseteq V \times K$ es la relación de etiquetado de las relaciones espaciales. Mediante esta definición pueden modelarse aristas paralelas representando distintas relaciones espaciales entre los pares de objetos de la base de datos, distinguiéndose por sus etiquetas.

Una vez obtenido el modelo gráfico de las relaciones espaciales entre los objetos de cada vecindario se buscan las asociaciones espaciales recordando que estas se definen como predicados compuestos conformados por varios predicados simples en los que por lo menos una relación espacial se ve involucrada, pudiendo ser el resto de los predicados relativos a los atributos no espaciales de los datos (Sección 2.1).

Para extraer las asociaciones en el espacio con alta tasa de ocurrencia se propone entonces la búsqueda de subgrafos frecuentes (sección 4.1.1) dentro de los grafos asociados a cada vecindario espacial, los cuales representan aquellas relaciones entre los tipos de datos espaciales que se presentan asiduamente en el modelo.

A modo de ejemplo, sea el subgrafo de la figura 4.11 parte de un modelo gráfico

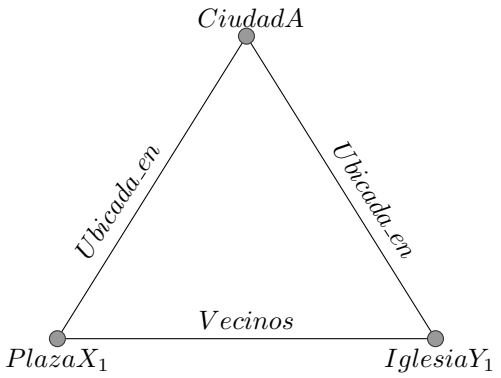


Figura 4.11: Ejemplo de relaciones entre objetos espaciales modeladas mediante un grafo.

más complejo, el siguiente predicado modela la interacción entre los elementos presentes en el mismo, cumpliendo con la definición de asociación espacial:

$$Ciudad(A) \wedge Plaza(X_1) \wedge Iglesia(X_2) \wedge$$
$$Ubicado\_en(X_1, A) \wedge Ubicado\_en(X_2, A) \wedge Vecinos(X_1, X_2)$$

Distintas herramientas pueden ser utilizadas para esta tarea, las cuales a su vez calculan la frecuencia de los subgrafos de formas dispares en función de parámetros como la recurrencia o el índice de compresión de información. Son ejemplos de las mismas: IncGM+, FSSG, SUBDUE, HSIGRAM, GREW, entre otros (Abdelhamid et al., 2017; Bianco, 2016; Kavitha et al., 2016).

Como resultado de esta fase se obtiene un conjunto de subgrafos frecuentes por cada vecindario que deberán ser analizados en la etapa siguiente con el fin de obtener conocimiento útil para la toma de decisiones, lo cual implica traducir los subgrafos a predicados, analizar la frecuencia de los mismos, determinar la relevancia de los mismos para el problema de negocio y calcular otras medidas de precisión en caso de ser necesario.

4.2.4. Proceso de descubrimiento de patrones de co-localización

El proceso de descubrimiento de patrones de co-localización (Fig. 4.12) aplica cuando el objetivo es determinar subconjuntos de tipos de objetos espaciales que fre-

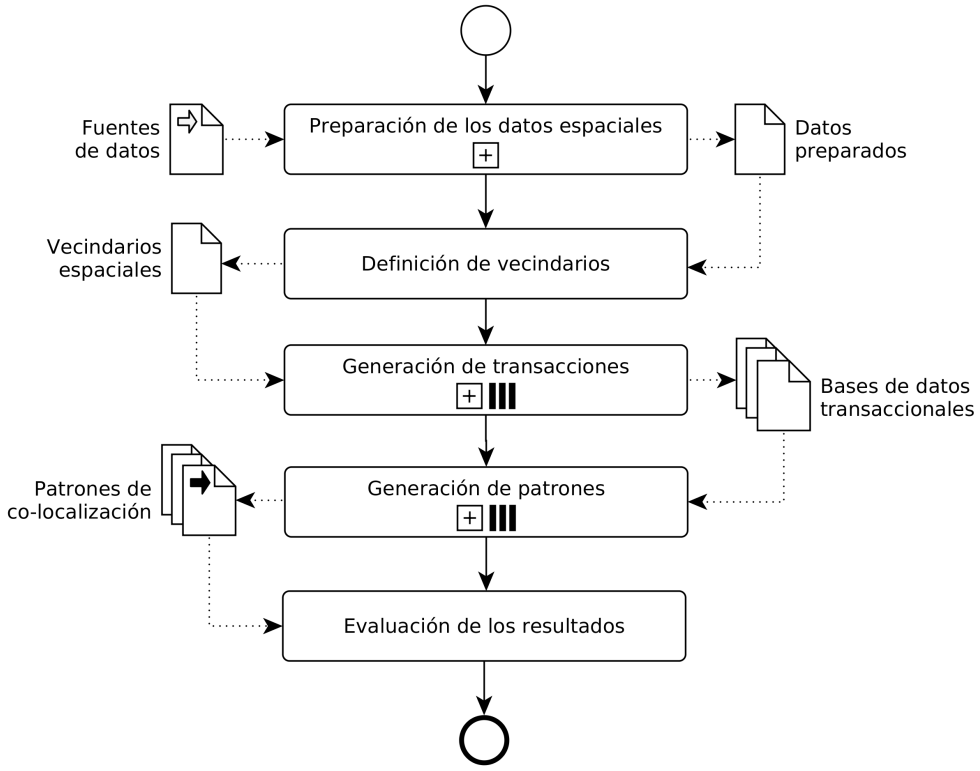


Figura 4.12: Proceso de descubrimiento de patrones de co-localización

cuentemente se encuentran vecinos entre sí, como pueden ser la determinación de especies simbióticas, condicionantes ambientales de enfermedades, asociaciones entre tipos de suelos y presencia de especies vegetales, entre otros.

Como se ha mencionado anteriormente, esta actividad puede verse como un tipo particular de descubrimiento de asociaciones espaciales, con la restricción de que, a diferencia de esta última, todas las instancias deben estar relacionadas entre sí conformando un clique, por lo que haremos uso de este concepto de la teoría de grafos para encontrar las regularidades.

El framework propuesto utiliza un enfoque transaccional para la generación de transacciones, lo que permite, al igual que en el proceso de descubrimiento de asociaciones espaciales, reutilizar los datos preprocesados en múltiples interacciones con los algoritmos de minería de datos sin tener que recalcular las relaciones espaciales en cada oportunidad.

En primer paso del proceso consiste, como hasta el momento, de la preparación de los datos para luego definir vecindarios espaciales de igual forma que en los procesos

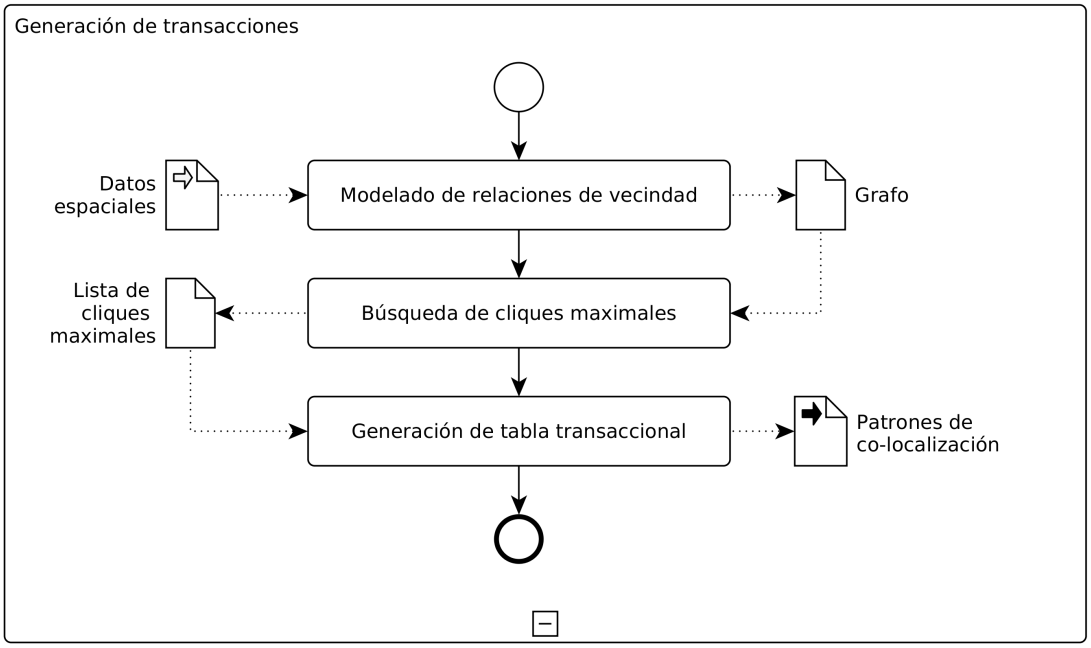


Figura 4.13: Subproceso de generación de transacciones para el descubrimiento de patrones de co-localización

anteriores. Luego, el proceso busca generar transacciones por cada uno de los vecindarios espaciales detectados. Para esto se aplica el subproceso de generación de transacciones a partir de las instancias de datos espaciales mediante estructuras de grafos, para luego obtener las regularidades ocultas en los mismos mediante un subproceso de generación de patrones de co-localización basado en aprendizaje supervisado o no supervisado, de acuerdo a los intereses de los tomadores de decisiones. A continuación se detallan estos dos subprocesos.

4.2.4.1. Generación de transacciones

El subproceso de generación de transacciones, como su nombre lo indica, busca la materialización de los vecindarios de puntos espaciales en una base de datos transaccional. Para esto, se hace uso de un grafo simple cuyos vértices representan objetos espaciales y cuyos aristas inciden sobre un par de puntos extremos si estos representan objetos espaciales que son vecinos entre sí, tal como se ejemplifica a continuación, siendo $Dist(x, y)$ una función de distancia entre dos objetos espaciales.

$$\forall x, y \in V, \{x, y\} \in A \iff Dist(x, y) \leq \delta \in \mathbb{R}$$

A partir de este grafo se identifican cliques maximales (Def. 4.12) haciendo uso de algoritmos especializados para tal propósito. Estos cliques maximales, entonces, conformarán una transacción de la base de datos transaccional en la que se garantiza que todos los elementos son vecinos entre sí (Kim et al., 2014). Esta transacción puede representarse mediante la tupla que se muestra a continuación:

$$T_i = (A = 1, B = 1, C = 0, D = 0, E = 1, F = 1)$$

Otros acercamientos dividen el espacio subyacente en grillas, cada una de las cuales se traduce en una transacción. Sin embargo, este acercamiento no garantiza que todos los elementos sean vecinos entre sí y puede ignorar relaciones de vecindad entre algunas instancias de datos.

El acercamiento basado en cliques máximos permite asegurar la completitud y la exactitud de los resultados obtenidos, según Kim et al. (2011, 2014).

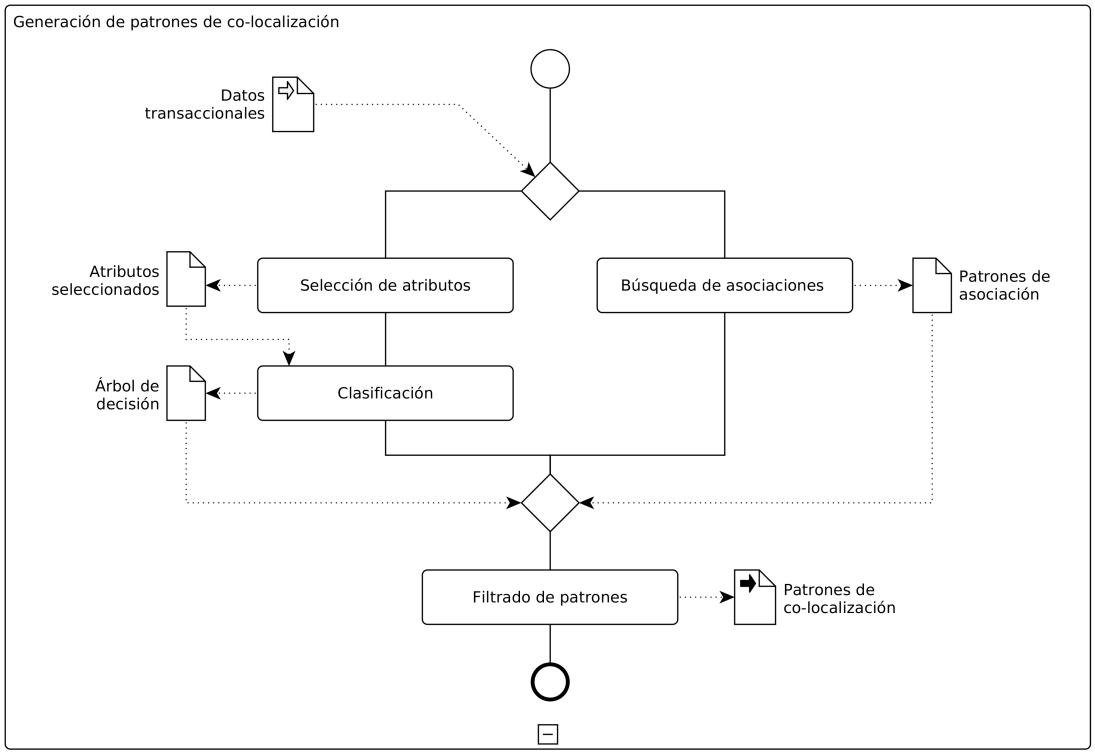


Figura 4.14: Subproceso de generación de patrones de co-localización a partir de datos transaccionales

4.2.4.2. Generación de patrones de co-localización

Utilizando el conjunto de datos transaccionales generado mediante el subproceso de generación de transacciones, se pueden utilizar dos estrategias para la obtención de patrones de co-localización.

En primer lugar, es factible utilizar algoritmos de búsqueda de patrones de asociación tradicionales, como los algoritmos *A-priori* (Agrawal et al., 1994) o FP-Growth (Han et al., 2000), sobre el conjunto de datos puesto que implícitamente se encuentran relacionados mediante la función de vecindad. De esta forma, los conjuntos de *items* frecuentes serán conjuntos de tipos de objetos espaciales que se encuentran frecuentemente vecinos entre sí.

Por otro lado, en el caso de que haya un tipo de evento espacial de particular interés para la inteligencia de negocio, es posible seleccionar dicho atributo como objetivo y obtener a partir de métodos de clasificación que otros objetos espaciales deben o no deben estar presente para que ocurra o no el primero.

En este caso se aporta más información para la toma de decisiones puesto que

la ausencia de un tipo de objeto espacial puede correlacionarse con la presencia o ausencia del tipo de objeto espacial de interés. No obstante, muchos patrones pueden perderse al considerar solo un atributo objetivo.

Por último, aquellos patrones obvios o triviales o aquellos que no cumplan con los requerimientos mínimos de confianza y calidad pueden ser descartados en la etapa de evaluación de resultados. Como salida de este proceso obtendremos un conjunto de patrones o reglas de co-localización de tipos de datos espacialmente referenciados.

4.3. Modelado difuso de relaciones espaciales

En una gran variedad de escenarios que involucran a datos espacialmente referenciados, resulta útil el modelado de las relaciones espaciales mediante la teoría de conjuntos difusos, desarrollada brevemente en la sección 4.1.2.

Esta idea se fundamenta en el hecho de que las relaciones espaciales no suelen ser rígidas. Por ejemplo, consideremos la relación rígida de vecindad entre dos puntos espaciales definida a como se muestra a continuación:

$$\forall x, y \in D, \mu_R(x,y) = \begin{cases} 1 & \text{si } Distancia(x,y) \leq \delta \\ 0 & \text{en caso contrario} \end{cases}, \text{con } \delta \in \mathbb{R}^+$$

Resulta intuitivo entender que, en un caso real, el limitar la distancia mínima a un valor determinado puede resultar en perdida de información relevante. Por ejemplo, si $\delta = 5$, estaríamos excluyendo de la relación a los pares de objetos que se encuentren a una distancia de 5.01 unidades, lo cual impactaría en el cálculo de las medidas de prevalencia del patrón, lo que puede ocasionar a su vez que el mismo sea ignorado.

Consecuentemente, el uso de la teoría de conjuntos difusos para el modelado de relaciones resulta conveniente. De esta forma, la relación de vecindad definida anteriormente puede reformularse de la siguiente manera, con $\delta_1, \delta_2, \gamma \in \mathbb{R}^+$:

$$\forall x, y \in D, \mu_{\tilde{R}}(x,y) = \begin{cases} 1 & \text{si } Distancia(x,y) \leq \delta_1 \\ ((\delta_1 - \delta_2)^{-1}(x - \delta_1) + 1)^{\gamma} & \text{si } \delta_1 < Distancia(x,y) \leq \delta_2 \\ 0 & \text{si } Distancia(x,y) > \delta_2 \end{cases}$$

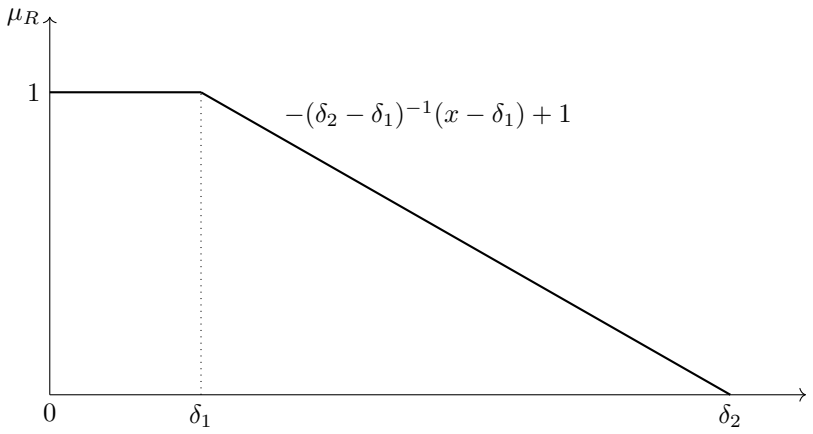


Figura 4.15: Ejemplo de definición de relación de vecindad difusa en función de la distancia entre dos elementos de la base de datos.

Considerando el valor del ponderador $\gamma = 1$, la relación de vecindad se puede representar gráficamente como se observa en la figura 4.15.

De igual manera pueden definirse las relaciones direccionales en función del ángulo que se forma entre dos puntos del plano. Por ejemplo, la relación Norte(A,B), leída como "A se encuentra al norte de B" se puede definir como sigue, quedando plasmada además en la figura 4.16.

$$\forall A, B \in D, \mu_{\tilde{Norte}}(A, B) = \begin{cases} 1 & \text{si } A_x = B_x \wedge A_y > B_y \\ |\frac{2}{\pi}\tan^{-1}(\frac{A_y - B_y}{A_x - B_x})| & \text{si } A_x \neq B_x \wedge A_y > B_y \\ 0 & \text{en caso contrario} \end{cases}$$

Esta definición se puede aplicar para otras direcciones mediante rotaciones de función y de las condiciones de cada una de las partes.

Asimismo, las relaciones topológicas pueden definirse de forma difusa si los objetos espaciales intervinientes no tienen fronteras rígidas. Por ejemplo, dado una zona de bosque, no es clara la frontera en donde esta zona termina, pero si puede considerarse distintas fronteras entre las que se apliquen funciones que graduen el grado de pertenencia a la región. De esta forma, relaciones como "A está dentro de B" se puede entender dependiendo de su posición con respecto a estas fronteras difusas, o la relación "A interseca con B" entre dos polígonos difusos puede modelarse a partir de la aplicación de la T-Norma entre los máximos valores de pertenencia que intervienen

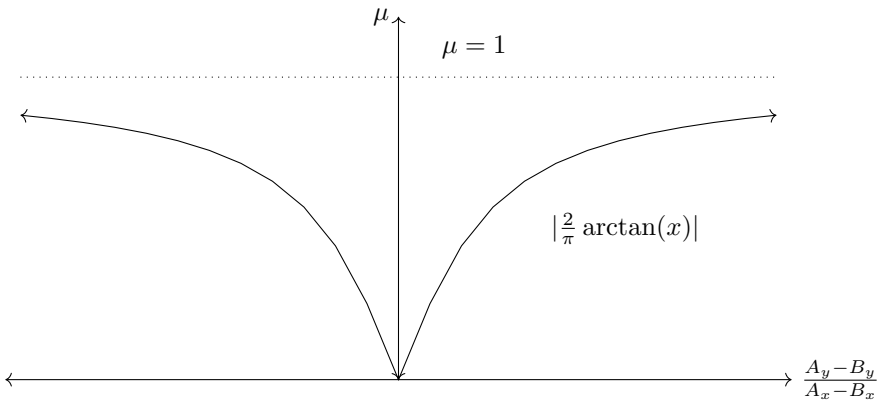


Figura 4.16: Función de pertenencia de la relación Norte(A,B) en función de la pendiente entre ambos puntos.

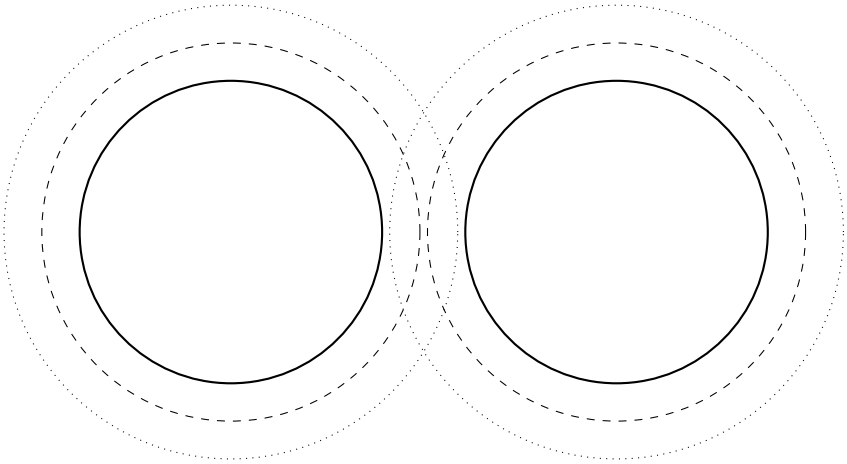


Figura 4.17: Ilustración que ejemplifica la intersección de dos regiones difusas. Cada grupo de círculos concéntricos representa una región cuyo valor de pertenencia transiciona entre las fronteras marcadas con línea recta, a trozos y punteada.

para cada región (Fig 4.17).

Estas propuestas de definición de relaciones espaciales son ejemplificaciones completamente arbitrarias. Las mismas deben definirse para los distintos tipos de objetos espaciales vectoriales evaluados teniendo en cuenta las características del dominio del problema.

Si bien las relaciones espaciales intervienen en todos los procesos de explotación de información propuestos, es en el proceso de descubrimiento de asociaciones espaciales donde múltiples relaciones se conjugan, debiendo determinar los grados de pertenencia de las relaciones involucradas en cada regularidad descubierta. Este detalle se desarrolla en el siguiente apartado.

4.3.1. Relaciones difusas en asociaciones espaciales

Como mencionamos anteriormente, el proceso de descubrimiento de asociaciones espaciales permite la obtención de proposiciones compuestas formadas por predicados entre los que se encuentra un subconjunto no vacío de los mismos que definen relaciones espaciales entre los objetos de la base de datos, o lo que es lo mismo, un predicado en el que se ven involucrados los atributos espaciales de los datos vectoriales considerados. Como ha sido mencionado previamente, estas relaciones poseen por lo general una naturaleza vaga, por lo que resulta apropiado hacer uso de la teoría de conjuntos difusos para modelarlas.

El proceso de descubrimiento de asociaciones espaciales utiliza teoría de grafos para el modelado de las relaciones y posteriormente el uso de minería de grafos a través de la búsqueda de subconjuntos frecuentes para hallar estas regularidades (ver sección 4.2.3). Este proceso puede ser adaptado para la utilización de relaciones espaciales difusas mediante una modificación en las actividades de modelado de relaciones espaciales y evaluación de resultados.

En primer lugar, como ya hemos mencionado en la sección 4.1.2, por cada uno de los vecindarios se extraen las relaciones espaciales de interés para el problema abordado, almacenando el grado de pertenencia de cada una de los elementos de la relación.

Posteriormente se modela mediante un grafo todos los elementos que pertenecen al soporte de cada una de estas relaciones espaciales, esto es, todos aquellos elementos o pares de objetos espaciales que se encuentren relacionados en algún grado mayor que cero. A partir de este grafo se extraen los subgrafos frecuentes tal como se realiza en el proceso con relaciones rígidas.

Luego, por cada patrón hallado se calcula el grado de pertenencia conjunto de cada instancia del mismo utilizando los grados de pertenencia de las instancias de las relaciones individuales que la componen. Para esto, por cada instancia se calcula la T-Norma de los grados de pertenencia, pudiendo variar esta según el dominio del problema.

En símbolos, sean $P(X) = p_1 \wedge p_2 \wedge \cdots \wedge p_m$ un predicado compuesto espacial, el conjunto de instancias de P, $V_P = \{\bar{X} | \mu_P(\bar{x}) > 0\}$, es el conjunto de todos los vectores de objetos espaciales $\bar{X} = (x_1, x_2, \cdots, x_k)$ que cumplen simultáneamente con P. Este grado de pertenencia se calcula mediante el uso de alguna t-norma $\top$ sobre los grados de pertenencia de los elementos de $\bar{X}$ a los predicados atómicos que lo componen:

$\mu_P(\bar{X}) = \top_{i=1..m} \mu_{p_i}(\bar{X})$.

A modo de ejemplo, supongamos el siguiente patrón de asociación, que se puede traducir como "Hay guaridas de leones en las llanuras que lindan con bosques":

$$Bosque(X) \wedge Llanura(Y) \wedge Guarida(Z) \wedge Intersecta(X, Y) \wedge Dentro(Z, X)$$

La tabla 4.1 muestra el grado de pertenencia de las distintas instancias del patrón P utilizando dos T-normas: la T-norma mínimo (Eq. 4.1) y la T-norma producto de Hamacher (Eq. 4.5 con $\lambda = 0$).

Como se puede observar, el grado de pertenencia del patrón utilizando $\top_{min}$ para las instancias 5 y 6 es igual a 0.80 en ambos casos, aún siendo mayor el grado de pertenencia del predicado "Intersecta" en el caso de la última instancia. Esta diferencia impacta en el valor de $\top_{H_0}$ reflejando de una forma más adecuada las diferencias de valores del predicado, otorgando un valor de pertenencia a la instancia 5 menor que el de la instancia 6. La diferencia entre estas dos funciones puede verse reflejada en la figura 4.5.

Asimismo, la T-norma de Hamacher es una función parametrizada, pudiendo adaptarse a las distintas circunstancias particulares del dominio del problema ajustando el parámetro λ de la función, tal como se observa en la formula 4.5 y en la figura 4.18.

Es posible a su vez calcular la T-norma de estos grados de pertenencia, con el fin de hallar un valor de pertenencia representativo del patrón en general, o bien usar descriptores estadísticos como el promedio o la desviación estándar para el mismo propósito.

Finalmente, esta misma estrategia se puede utilizar en los demás procesos de explotación de información en los cuales intervienen relaciones espaciales explícitas, sin embargo resulta más relevante y claro en el proceso de descubrimiento de asociaciones espaciales.

Tabla 4.1: Ejemplo de cálculo del grado de pertenencia de predicado compuesto por tres predicados rígidos y dos predicados espaciales difusos para seis instancias utilizando $\top_{min}$ y $\top_{H_0}$

Inst.	**Bosque**	**Llanura**	**Guarida**	**Intersecta**	**Dentro**	$\top_{min}(\bar{x}_i)$	$\top_{H_0}(\bar{x}_i)$
1	1.00	1.00	1.00	0.70	0.90	0.70	0.65
2	1.00	1.00	1.00	0.54	0.87	0.54	0.49
3	1.00	1.00	1.00	0.40	0.70	0.40	0.34
4	1.00	1.00	1.00	0.60	0.20	0.20	0.17
5	1.00	1.00	1.00	0.80	0.90	0.80	0.73
6	1.00	1.00	1.00	0.80	1.00	0.80	0.80
					$\top$	0.2	0.099
					Promedio	0.57	0.53
					SD	0.24	0.24

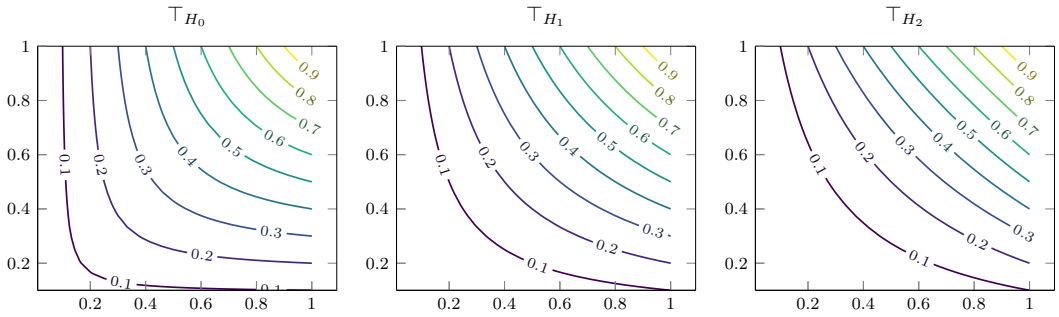


Figura 4.18: Comparativa de instancias de la familia de T-Norma de Hamacher para $\lambda = 0$ (producto de Hamacher), $\lambda = 1$ (T-norma producto) y $\lambda = 2$, respectivamente.

CAPÍTULO 5

VALIDACIÓN

> "Never trust the storyteller. Only trust the story."
>
> *Neil Gaiman, The Hunt*

Una vez propuestos los artefactos solución es necesario determinar si la interacción entre estos artefactos y el contexto permiten solucionar el problema que da origen al objetivo planteado (sección 3.1). Para esto, el framework de la ciencia de diseño menciona dos actividades: la demostración del artefacto y la evaluación del mismo. En este capítulo llevaremos adelante ambas actividades por estar íntimamente relacionadas.

En primer lugar, la demostración del artefacto, según Johannesson and Perjons (2014) permite ilustrar el uso del artefacto en un caso particular, dando así pruebas de su factibilidad. En otras palabras, la demostración del artefacto permite responder la pregunta: *¿Cómo puede ser usado el artefacto para resolver el problema explicado para un caso particular?*.

Para esto, por cada uno de los procesos de explotación de información diseñados se selecciona un caso de demostración y se describe cómo se instancia cada una de las actividades que forman parte de los mismos y de qué forma se puede obtener conocimiento para la toma de decisiones (Fig. 5.1). Por otro lado, más allá de que la demostración puede considerarse como una forma débil de evaluación, esta sirve además para ejemplificar y bajar el nivel de abstracción de los artefactos construidos de forma tal que estos ejemplos puedan ser utilizados seguidamente en la fase de evaluación.

Por el otro lado, la evaluación permite determinar qué tan bien el artefacto permite resolver el problema y cumplir con los requerimientos (Fig. 5.2). Permite responder a

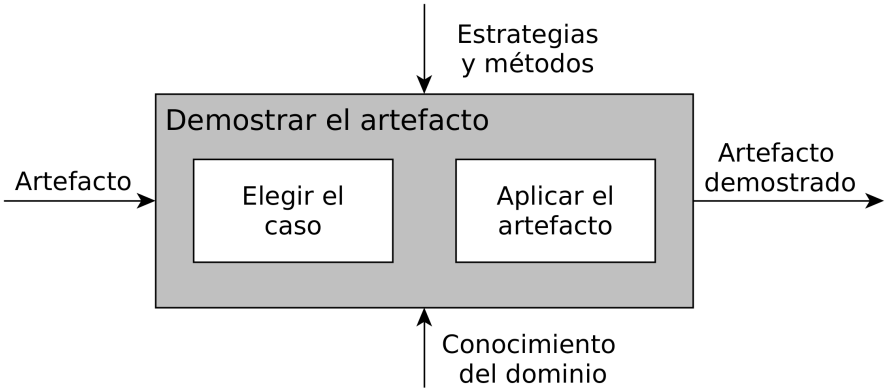


Figura 5.1: Actividades del framework de ciencia del diseño para demostración del artefacto. Traducido y adaptado de (Johannesson, 2014).

la pregunta: *¿Qué tan bien el artefacto soluciona el problema explicado y cumple con los requerimientos?* (Johannesson and Perjons, 2014). En el contexto de este trabajo, la actividad de evaluación posee tres objetivos:(1) evaluar la utilidad del artefacto; (2) determinar el cumplimiento de los requerimientos; (3) Identificar oportunidades de mejora.

Para esto, según Wieringa (2014), existen múltiples métodos para el estudio de estos modelos de validación, entre los que se destaca la validación por la opinión de expertos por su sencillez. Asimismo, según el esquema conceptual de Hevner et al. (2004) para la evaluación de artefactos (resumidos en la tabla 5.1), resulta adecuado un estudio descriptivo para el tipo de artefacto propuesto en esta tesis doctoral. Estos procesos proponen una estructura guía para la obtención de conocimiento, ordenando y estandarizando el proceso, por lo que resultan relevantes en relación con su interacción con sus usuarios.

Según Wieringa (2014), la validación por juicio de expertos consiste en la exposición del diseño de los artefactos a un panel de expertos. Estos expertos imaginan cómo el artefacto interactúa con situaciones del contexto hipotéticas y predicen qué efectos piensan que pueden tener. Este acercamiento es muy similar a la técnica de inspección de código, en el cual un número de ingenieros de software leen el código escrito por alguien más para encontrar errores o problemas potenciales.

El objetivo de la validación de expertos no consiste en realizar una encuesta de las opiniones de los expertos sino que, por el contrario, estos son usados como instru-

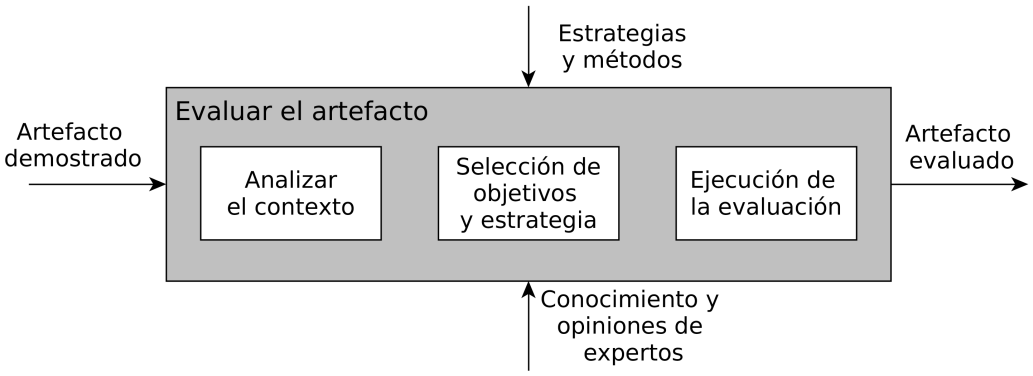


Figura 5.2: Actividades del framework de ciencia del diseño para evaluación del artefacto. Traducido y adaptado de (Johannesson, 2014).

mentos de observación, por lo que es necesario que ellos posean el nivel de experticia adecuado para entender el artefacto, imaginar problemas realistas y generar predicciones confiables sobre el efecto de los mismos en el contexto. En este caso, las opiniones negativas son las más valoradas para poder determinar puntos de mejora.

No obstante, si bien los expertos son capaces de imaginar la interacción artefacto-contexto, los casos de demostración de funcionamiento del artefacto sirven para ilustrar de qué manera estos pueden ser usados para la extracción de conocimiento en dominios específicos, ilustrar de qué forma estos procesos pueden ser implementados. Asimismo, las demostraciones sirven también de ejemplo guía para el panel de expertos, equiparable al uso de escenarios (Hevner et al., 2004), a partir de los cuales se podrán determinar la utilidad de los mismos, nuevas formas de implementación y cuestiones como la complejidad para el entendimiento de los resultados.

Los casos de demostración se encuentran ordenados como sigue: en la sección 5.1 se detallan las implementaciones del proceso de descubrimiento de grupos espaciales; en la sección 5.2 las implementaciones del proceso de descubrimiento de anomalías espaciales; luego, en la sección 5.3 se describen las implementaciones del proceso de descubrimiento de asociaciones espaciales; y, por último, en la sección 5.4 se lee lo referente a la implementación del proceso de descubrimiento de patrones de co-localización.

Las opiniones de los expertos en cuanto a las dimensiones evaluadas de los procesos de explotación de información serán agregadas mediante el uso de teoría de

Tabla 5.1: Esquema de clasificación de métodos de evaluación de diseño de artefactos según Hevner et al. (2004).

Tipo	**Método**	**Descripción**
Observacional	Estudio de Caso	Estudio del artefacto en profundidad en su contexto de negocio
	Estudio de Campo	Monitorización del uso de un artefacto en múltiples proyectos.
Analítico	Análisis Estadístico	Examinar la estructura del artefacto según sus cualidades estáticas (Ej. complejidad)
	Análisis de Arquitectura	Estudio de la adecuación del artefacto en arquitecturas técnicas de sistemas de información.
	Optimización	Demuestra las propiedades inherentemente óptimas del artefacto o provee los límites entre los cuales el artefacto demuestra ser óptimo.
	Análisis Dinámico	Estudia las características dinámicas de un artefacto (Ej. eficiencia).
Experimental	Experimento controlado	Estudia el artefacto en un contexto controlado.
	Simulación	Ejecuta el artefacto con datos artificiales.
Testing	Funcional (caja negra)	Ejecuta las interaces del artefacto para determinar fallas e identificar defectos de funcionamiento.
	Estructural (caja blanca)	Ejecuta *testing* de cobertura para la determinación de métricas sobre la implementación del artefacto.
Descriptivo	Argumento informado	Uso de información del estado del arte para construir un argumento convincente sobre la utilidad del artefacto.
	Escenarios	Construye escenarios detallados alrededor del artefacto para demostrar su utilidad.

conjuntos difusos, cuestión que se encuentra plasmada en la sección 5.5. Por último, en la sección 5.6 se resume la información obtenida en este capítulo y se analizan en profundidad las opiniones de los expertos para con los artefactos propuestos.

5.1. Descubrimiento de grupos espaciales

En este apartado se contextualiza el problema abordado para la validación del proceso de descubrimiento de grupos espaciales (Sección 5.1.1), se detallan los datos utilizados (Sección 5.1.2), se describe el objetivo de minería de datos (Sección 5.1.3), se detallan los pasos y herramientas utilizadas en la implementación de las actividades del proceso (Sección 5.1.4), y se describe la evaluación de los resultados y la evaluación realizada por el experto (Sección 5.1.5).

5.1.1. Contexto del problema

AirBnB (http://airbnb.com/) es una empresa fundada en el año 2008 por Brian Chesky, Joe Gebbia y Nate Blecharczyk, que brinda una plataforma en línea para la oferta de alojamientos a particulares y turísticos. A través de la misma los anfitriones pueden publicitar y contratar el arriendo de sus propiedades con sus huéspedes; anfitriones y huéspedes pueden valorarse mutuamente, como referencia para futuros usuarios (McCann, 2015).

Esta compañía posee en el año 2020 más de 30 oficinas alrededor del mundo y opera en 191 países brindando servicio a más de 150 millones de usuarios que pueden optar entre más de 5 millones de opciones para su alojamiento (Needed, 2020).

5.1.2. Datos utilizados

Los datos utilizados (Rottoli, 2020a) contienen información sobre propiedades de AirBnB en el área interior de Londres agregada a nivel “Middle Layer Super Output Area” (MSOA). El dataset fue originalmente compilado por Dani Arribas-Bel usando datos originales de AirBnB (http://insideairbnb.com/). Los datos utilizados poseen licencia CC0 1.0 Universal y se encuentran disponibles en un único GeoJson (Arribas-Bel, 2020).

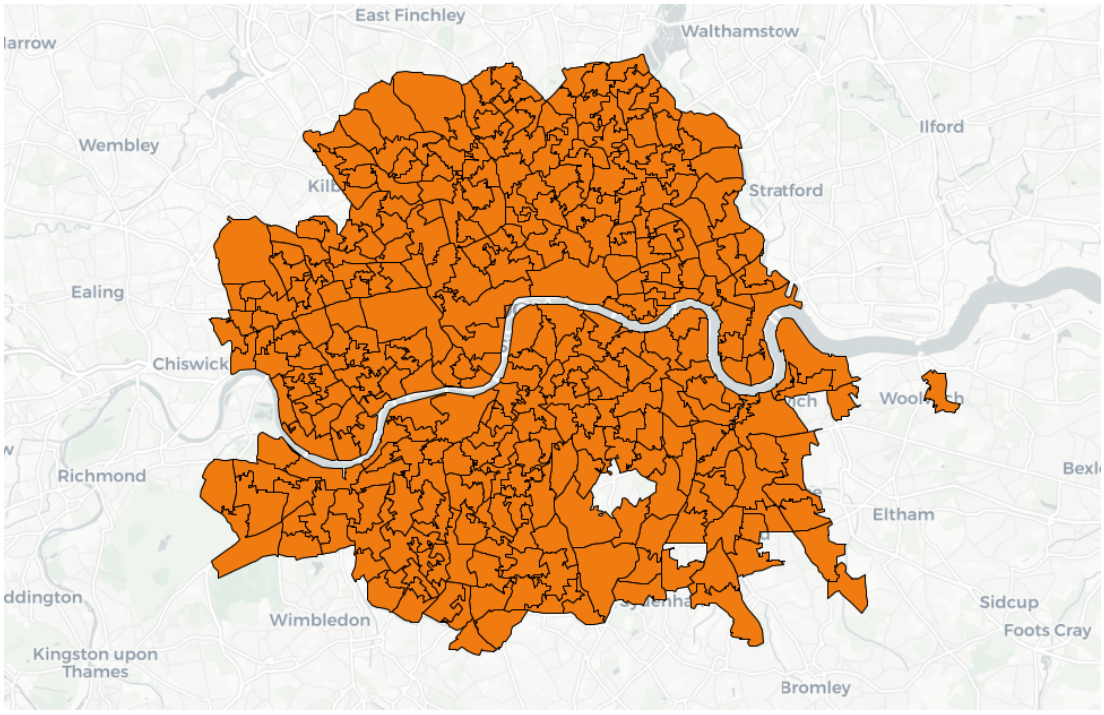


Figura 5.3: Área interior de Londres a nivel MSOA. Polígonos (en naranja) utilizados en la validación del proceso de descubrimiento de grupos espaciales.

El conjunto de datos crudo consta de 320 polígonos, que se pueden observar en la figura 5.3 descriptos por 15 atributos, los cuales son detallados en la tabla 5.2, con valores promedios descriptivos de las propiedades que se encuentran en cada una de las regiones consideradas y valores promedio de la opinión volcada por los inquilinos de dichas propiedades.

Con excepción del atributo *id*, todos los atributos enumerados son atributos numéricos reales cuya distribución de valores se encuentra detallada en la tabla 5.3. Se evidencian posibles valores anómalos que deberán ser considerados en la preparación de los datos (Fig. 5.4).

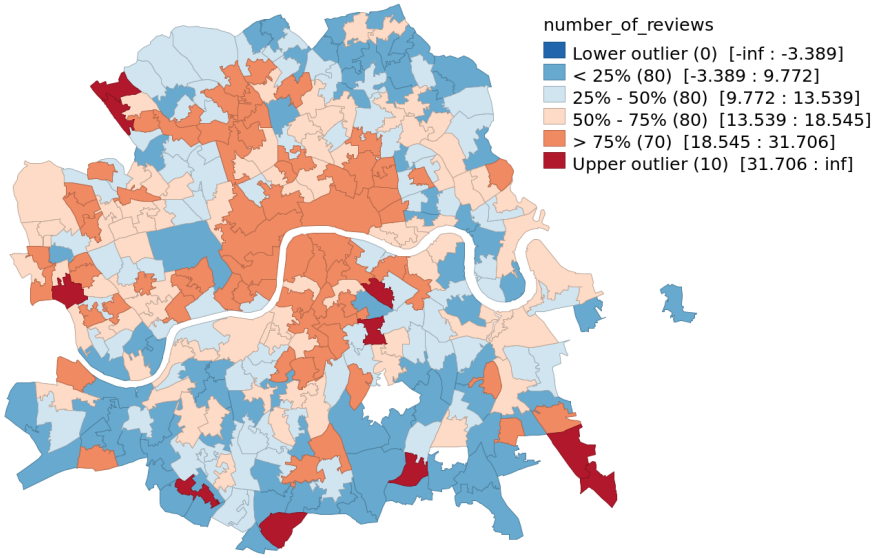
5.1.3. Objetivos

En el contexto de la presente validación se plantea como objetivo determinar particiones del área interior de Londres según las calificaciones obtenidas en AirBnB por los clientes.

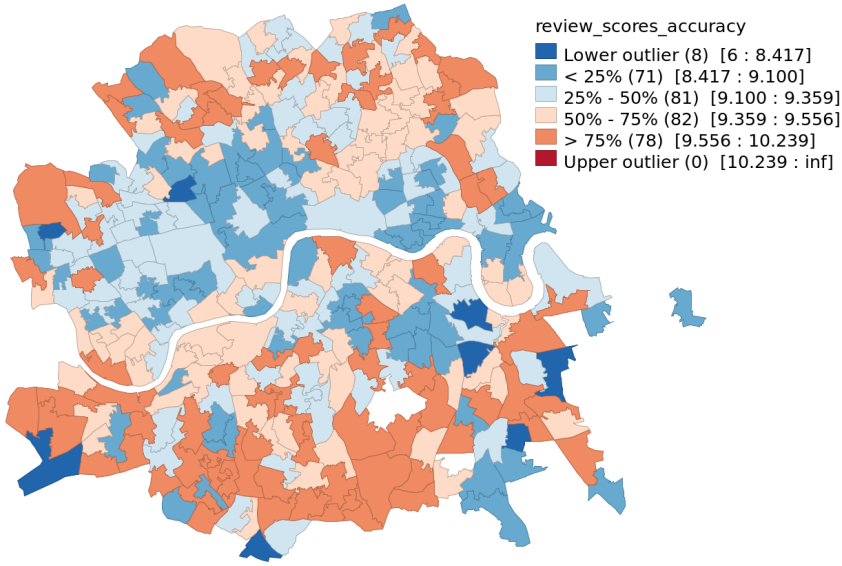
Para esto, se realizará una implementación del proceso de explotación de infor-

Tabla 5.2: Descripción de los atributos del dataset utilizado en la validación del proceso de descubrimiento de grupos espaciales.

Atributo	**Descripción**
Id	Identificador único del área MSOA.
Accommodates	Capacidad promedio de las propiedades en el MSOA.
Bathrooms	Número promedio de baños en las propiedades en el MSOA.
Bedrooms	Número promedio de dormitorios en las propiedades en el MSOA.
Beds	Número promedio de camas en las propiedades en el MSOA.
Number_of_reviews	Número promedio de revisiones recibidas para las propiedades en el MSOA.
Reviews_per_month	Número promedio de revisiones por mes recibidas para las propiedades en el MSOA.
Review_scores_ratings	Valoración promedio recibido para las propiedades en el MSOA.
Review_scores_accuracy	Valoración promedio de precisión recibido para las propiedades en el MSOA.
Review_scores_cleanliness	Valoración promedio de limpieza recibido para las propiedades en el MSOA.
Review_scores_checkin	Valoración promedio del proceso de *checkin* recibido para las propiedades en el MSOA.
Review_scores_communication	Valoración promedio de la comunicación recibido para las propiedades en el MSOA.
Review_scores_location	Valoración promedio de la ubicación de las propiedades en el MSOA.
Review_scores_value	Valoración promedio general recibido para las propiedades en el MSOA.
property_count	Cantidad de propiedades listadas en AirBnB que se encuentran en el MSOA.



(a)



(b)

Figura 5.4: Mapa que ilustra la ubicación de las regiones anómalas con respecto a los atributos (a) *number_of_reviews* y (b) *review_scores_accuracy*.

Tabla 5.3: Estadísticas descriptivas de los atributos del conjunto de datos utilizado para la validación del proceso de descubrimiento de grupos espaciales.

	$\bar{X}$	**S.D.**	**Mín**	**25 %**	**50 %**	**75 %**	**Máx**
Accommodates	3,85	0,96	1,00	3,15	3,50	3,91	11,0
Bathrooms	1,27	0,42	0,85	1,13	1,22	1,33	8,00
Bedrooms	1,49	0,39	0,50	1,28	1,44	1,60	4,00
Beds	2,00	0,85	1,00	1,67	1,88	2,11	11,00
Number_of_reviews	15,01	9,15	1,00	9,78	13,54	18,52	71,00
Reviews_per_month	1,36	0,54	0,14	1,00	1,29	1,64	3,85
Review_scores_rating	91,01	5,08	40,00	89,31	91,50	93,67	100,00
Review_scores_accuracy	9,32	0,40	6,00	9,10	9,36	9,56	10,00
Review_scores_cleanliness	9,14	0,55	2,00	8,94	9,15	9,40	10,00
Review_scores_checkin	9,52	0,45	4,00	9,38	9,55	9,76	10,00
Review_scores_comm...	9,59	0,54	4,00	9,47	9,66	9,83	10,00
Review_Scores_location	9,09	0,54	7,00	8,80	9,12	9,50	10,00
Review_scores_value	9,01	0,50	4,00	8,82	9,00	9,25	10,00
Property_count	24,47	26,09	1,00	8,00	17,00	33,00	203,00

mación propuesto haciendo uso de herramientas de regionalización para el particionamiento del área geográfica evaluada.

5.1.4. Implementación del proceso de explotación de información

El proceso para el descubrimiento de grupos espaciales es encuadrado en la metodología CRISP-DM tal como se observa en la figura 5.5. La actividad de preparación de los datos del proceso de explotación de información se implementa como actividades de la fase homónima del marco metodológico, mientras que los subprocesos de agrupamiento de los datos espaciales y de descripción de los grupos se realizan en el contexto de la fase de modelado.

A continuación se detallan estas fases en sus respectivas subsecciones.

5.1.4.1. Preparación de los datos

Primeramente, a partir del conjunto de datos previamente mencionado en formato GeoJSON, se obtuvo el conjunto de atributos no espaciales en formato CSV mediante el uso de la aplicación QGis (Qgis, 2020) para ser utilizado como entrada de la aplicación para minería de datos Rapidminer (RapidMiner, 2020).

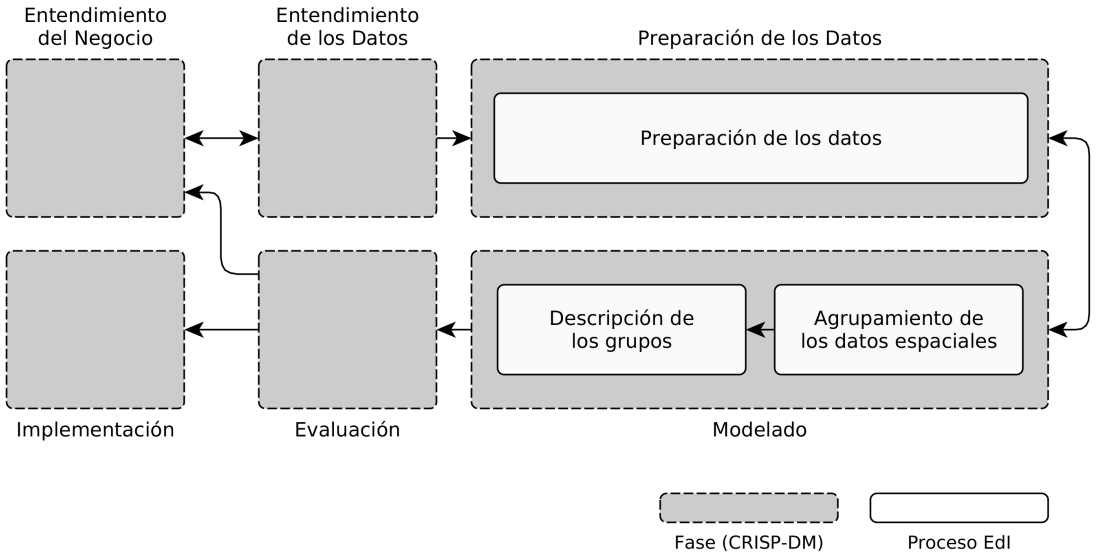


Figura 5.5: Implementación del proceso de descubrimiento de grupos espaciales dentro de la metodología CRISP-DM.

En esta última aplicación se seleccionaron solo aquellos atributos no espaciales relevantes para con los objetivos planteados anterioridad según el criterio del analista, siendo estos *number_of_reviews*, *reviews_per_month*, *review_scores_ratings*, *review_scores_ accuracy*, *review_scores_cleanliness*, *review_scores_checkin*, *review_scores_communication*, *review_scores_location*, *review_scores_value* y *property_count*.

Estas variables fueron normalizadas en el intervalo [0;1] aplicando al fórmula de la ecuación 5.1, donde $\bar{X}$ es el atributo a normalizar y X_i, con $i = 1..|\bar{X}|$, es valor del atributo en la fila i.

$$N(X_i) = \frac{X_i - Min(\bar{X})}{Max(\bar{X}) - Min(\bar{X})} \tag{5.1}$$

Esta detección de anomalías busca eliminar las anomalías globales de la base de datos y fue realizada mediante la ejecución del algoritmo LOF (Breunig et al., 2000) sobre los atributos seleccionados y normalizados. Se seleccionó el umbral de LOF para la selección de anomalías en valor = 2, tal como sugiere (Kuna et al., 2009), de tal forma que cualquier tupla con un valor superior a este umbral es considerada un *outlier*. De esta forma fueron identificadas las tuplas indicadas en la tabla 5.4, para luego ser eliminadas del mapa original utilizando la herramienta QGis.

Tabla 5.4: Anomalías globales detectadas en las bases de datos utilizada para la validación del proceso de descubrimiento de grupos espaciales.

Fila	ID	LOF
160	E02000652	5.726
187	E02000685	3.972
182	E02000677	2.590
298	E02000972	2.564
158	E02000650	2.357
6	E02000170	2.278
186	E02000683	2.174
31	E02000326	2.023
57	E02000371	2.005

5.1.4.2. Agrupamiento de datos espaciales

Los datos preparados en la actividad anterior fueron utilizados como entrada para el subproceso de agrupamiento de datos espaciales, el cual fue implementado mediante la aplicación GeoDa (Anselin et al., 2020), *software* de análisis de datos espaciales de código abierto.

En esta aplicación se selecciona la herramienta de agrupamiento REDCAP utilizando la estrategia *Full-Order Average Linkage* (Guo, 2008) y los 6 vecinos más cercanos de cada región administrativa como el criterio de contigüidad. No se utilizó la contigüidad topológica por existir regiones no conexas en el mapa. Los atributos fueron estandarizados (Z) y se seleccionó la cantidad de *clusters* en 4.

Tras la ejecución, se obtuvieron 4 grupos distribuidos tal como se observa en la figura 5.6 y cuyas métricas se detallan en la tabla 5.5: grupo 1 con 146 elementos, grupo 2 con 113 elementos, grupo 3 con 42 elementos y grupo 4 con 10 elementos.

Estos *clusters* identificados fueron integrados a la base de datos mediante la misma aplicación utilizada para generarlos agregando un atributo categórico denominado "CL".

5.1.4.3. Descripción de los grupos

Una vez identificadas las regiones, estas fueron utilizadas como entrada del proceso de descripción de grupos. Para esto, se utilizó Rapidminer seleccionado el atributo "CL" anteriormente mencionado como atributo objetivo estableciendo su rol como *la-*

Tabla 5.5: Métricas de la ejecución del algoritmo de regionalización sobre los datos utilizados para la validación del proceso de descubrimiento de grupos espaciales.

Métrica	**Valor**
Suma de cuadrados total	3100
Suma de cuadrados interna del grupo 1	969.068
Suma de cuadrados interna del grupo 2	785.08
Suma de cuadrados interna del grupo 3	584.283
Suma de cuadrados interna del grupo 4	50.8044
Suma de cuadrados interna total	2389.24
Suma de cuadrados entre grupos	710.764
Relación entre suma de cuadrados	0.229279

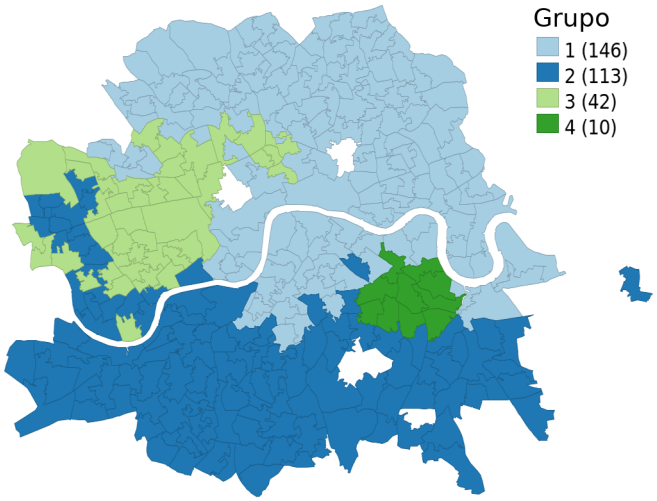


Figura 5.6: Grupos generados mediante regionalización sobre los datos de AirBnB

Tabla 5.6: Parámetros del algoritmo TDIDT utilizado sobre la regionalización de los datos de validación del proceso de descubrimiento de grupos espaciales.

Parámetro	**Valor**
Criterio de selección	Ganancia de información
Máxima profundidad	5
Confianza mínima	0.1
Pre-poda	Si
Poda	Si
Mínima ganancia	0.01
Mínimo tamaño de la hoja	5
Mínimo tamaño para dividir	5
Cantidad de pre-podas	3

bel y los atributos utilizados para la construcción de los grupos como atributos de entrada.

Se implementó un proceso de validación cruzada con 10 *folds* sobre los datos ingresados para obtener métricas sobre la clasificación de los datos utilizando distintas configuraciones del algoritmo de clasificación basado en árboles de decisión C4.5.

Los resultados mostrados a continuación (Fig. 5.7) corresponden a la clasificación con mayor valor de precisión obtenida sobre los grupos, 57.22 % +/- 8.55 % (micro promedio: 57.23 %), correspondiente a los parámetros indicados en la tabla 5.6.

Como se puede observar, la clasificación permite descubrir adecuadamente las regiones 1 y 2, mientras que las regiones 3 y 4 poseen un menor grado de precisión, por lo que sería adecuado explorar los centroides estadísticos de las mismas para poder obtener mayor información relevante.

Las reglas obtenidas para cada región son las siguientes:

1. Región 1:
   - `SI property_count > 24 Y review_scores_location ≤ 9.554 ENTONCES Región 1`
   - `SI property_count ≤ 24 Y review_scores_checkin > 9.323 Y reviews_per_month > 1.420 ENTONCES Región 1`
   - `SI 11 < property_count ≤ 24 Y review_scores_checkin ≤ 9.323 Y review_scores_cleanliness ≤ 8.846 ENTONCES Región 1`

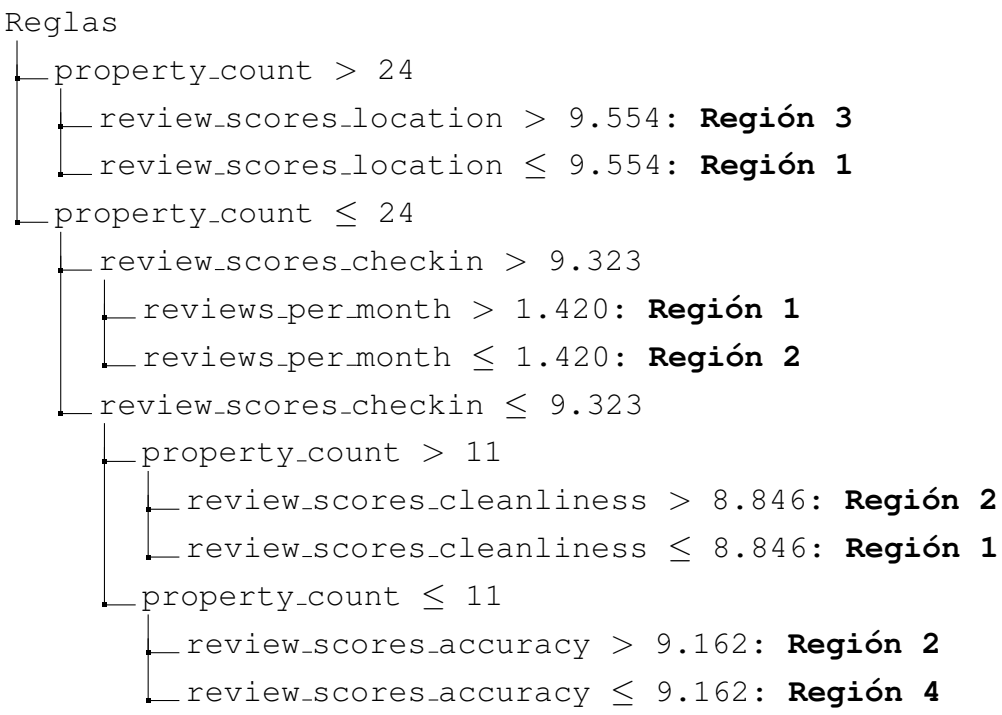


Figura 5.7: Árbol de decisión para las regiones obtenidas sobre los datos de AirBnB

2. Región 2:

   - SI property_count ≤ 24 Y review_scores_checkin > 9.323 Y reviews_per_month ≤ 1.420 ENTONCES Región 2
   - SI 11 < property_count ≤ 24 Y review_scores_checkin ≤ 9.323 Y review_scores_cleanliness > 8.846 ENTONCES Región 2
   - SI property_count ≤ 11 Y review_scores_checkin ≤ 9.323 Y review_scores_accuracy > 9.162 ENTONCES Región 2

3. Región 3:

   - SI property_count > 24 Y review_scores_location > 9.554 ENTONCES Región 3

4. Región 4

   - SI property_count ≤ 11 Y review_scores_checkin ≤ 9.323 Y review_scores_accuracy ≤ 9.162 ENTONCES Región 4

Tabla 5.7: Matriz de confusión de la clasificación realizada para el proceso de descubrimiento de grupos espaciales.

	True 1	True 3	True 2	True 4	Precisión
Pred. 1	84	22	34	4	58.33 %
Pred. 3	13	19	0	0	59.38 %
Pred. 2	45	1	72	3	59.50 %
Pred. 4	4	0	7	3	21.43 %
Recall	57.53 %	45.24 %	63.72 %	30.00 %	

5.1.5. Evaluación de los resultados

Del análisis de las reglas y regiones obtenidas mediante el proceso de descubrimiento de grupos espaciales para la regionalización de los datos de AirBnB detallados con anterioridad, se pueden extraer las siguientes conclusiones:

La región 1 se encuentra compuesta por 60 secciones con una alta cantidad de propiedades (más de 60 por sobre el promedio del atributo) pero con una evaluación de la localización menor que no llega a ubicarse entre el 25 % con mejores puntuaciones según este criterio (Tabla 5.3). En cambio, las secciones de la región 3 poseen en 20 casos una localización evaluada por encima de los 9.5 puntos por sobre 10 y una cantidad de oferta similar.

Por otro lado, se evidencian 67 secciones de la región 1 que, si bien tienen una menor cantidad de opciones en propiedades, poseen una puntuación del proceso de *checkin* por arriba de los 9.3 puntos, correspondiente al primer cuartil de los datos. Existen además 77 secciones de la región 2 que comparten estas características pero poseen una demanda mensual inferior.

En la región 1 además se encuentran casos de nueve secciones que poseen entre 11 y 24 propiedades, menos calidad en el proceso de *checkin* y una pobre puntuación de evaluación de la limpieza.

La región 2 por otro lado posee secciones con una cantidad de propiedades por debajo del promedio. 77 de estas secciones posen una valoración promedio del proceso de *checkin* por encima del primer cuartil, pero con una baja cantidad de revisiones por mes. Este comportamiento se comparte con ciertas secciones de la región 1, diferenciándose por estas últimas ser más concurridas. Otras 6 regiones poseen un comportamiento distinto al anteriormente mencionado pero con baja certeza.

Luego, la región 3 posee una cantidad de propiedades por sobre el promedio y una

valoración de sus ubicaciones ubicada dentro del 25 % mejor valoradas.

Por último, la región 4, posee una cantidad de propiedades inferior a 11, una baja apreciación del proceso de *checkin* y una valoración de la precisión de las valoraciones baja.

En conclusión, la región 3 posee una gran cantidad de oferta y bien posicionadas geográficamente, la región 4 posee una oferta pequeña y menor calidad en su proceso de checkin y la precisión de las revisiones, la región 2 se caracteriza por tener una oferta inferior al promedio, pocas revisiones mensuales pero una recepción bien puntuada, y por último la región 1 está compuesta por secciones con alta cantidad de propiedades pero no evaluadas de forma favorecedora en cuanto a posición, y por secciones con menos cantidad de oferta pero muy concurridas y bien criticadas.

Es necesario en este punto aclarar que si bien la precisión de la clasificación de la región 4 es baja, aún resulta útil para ilustrar el funcionamiento del proceso de explotación de información propuesto. Los valores de las métricas asociadas a la clasificación dependerán en cada caso de la selección de parámetros tanto para el agrupamiento como de los algoritmos de aprendizaje supervisado, así como también de las características del conjunto de datos.

5.2. Descubrimiento de anomalías espaciales

En esta sección se presenta el desarrollo de la validación del proceso de descubrimiento de anomalías espaciales. Para esto se contextualiza el problema de negocio abordado (Sección 5.2.1), se presentan los datos utilizados (Sección 5.2.2), se describen los objetivos de minería de datos (Sección 5.2.3), se detalla la implementación del proceso haciendo uso de herramientas varias (Sección 5.2.4), y por último se evalúan los resultados apelando a la valoración de un experto en el tema (Sección 5.2.5).

5.2.1. Contexto del problema

Los Estados Unidos de América, Estados Unidos o USA por sus siglas en inglés (*United States of America*), es un país soberano compuesto por cincuenta estados y un distrito federal. La mayor parte del país se ubica en el medio de América del Norte, limita con Canadá al norte y con México al sur. El estado de Alaska está en el noroeste

del continente, limita con Canadá al este, separado de Rusia al oeste por el estrecho de Bering. El estado de Hawai es un archipiélago en el océano Pacífico y es el único de sus estados que no se encuentra en América. El país posee en el mar Caribe y en el Pacífico varios territorios no incorporados (Brittingham and de la Cruz, 2004).

Estados Unidos posee alrededor de 325 millones de habitantes. Esto la convierte en la tercera nación más poblada en el mundo, después de China y la India. Su tasa de crecimiento demográfico es de 0,98 %, significativamente más alto que los de Europa occidental, Japón y Corea del Sur (Adams and Strother-Adams, 2001). Además, Estados Unidos es una de las naciones con más diversidad étnica y cultural, producto de la inmigración a gran escala teniendo en total treinta y un diferentes grupos étnicos que cuentan con más de un millón de representantes (U.S. Census Bureau, 2020b).

Los estados de Estados Unidos se dividen en niveles locales denominados condados (*county*) creados como una subdivisión de un estado por el gobierno estatal o por el federal o por el gobierno territorial. Este término se utiliza en 48 de los 50 estados estadounidenses; Luisiana utiliza el término "*parish*" (parroquia) y Alaska "borough". Existen 3098 condados actualmente en este país. En el censo estadounidense del 2000 solo el 16,7 % de condados o equivalentes tenía más de 100.000 habitantes. (Census.gov, 2013).

5.2.2. Datos utilizados

La base de datos utilizada se encuentra conformada por datos demográficos de los condados de Estados Unidos correspondientes al año 2017 recopilados por la oficina de censo del gobierno estatal. Estos datos muestran el número de personas que se encuentran clasificadas en franjas etarias pre-establecidas y en subgrupos étnicos, indicando esta totalidad en las columnas referidas a tales clasificaciones.

Se cuenta con una totalidad de 3221 registros, incluyendo secciones administrativas de Puerto Rico y Alaska y una única tupla correspondiente a los datos de la totalidad del país. Estos elementos de datos se encuentran descriptos por 358 atributos incluyendo identificadores, que no serán descriptos en el presente documento debido a tal extensión. Para más información los datos y sus descripciones se encuentran disponibles en (Rottoli, 2020d)

Tabla 5.8: Descripción de los atributos del dataset utilizados en la validación del proceso para descubrimiento de anomalías espaciales.

Atributo	Nuevo nombre	Descripción
GEO_ID	Id	Identificador del condado
-	FIPS	Código FIPS de identificación del condado.
NAME	Geographic Area Name	Nombre del área geográfica
DP05_0004E	Total Sex ratio	Población masculina por cada 100 de sexo femenino
DP05_0019E	Population under 18y	Población estimada menor de edad
DP05_0021E	Population over 17y	Población estimada mayor de edad
DP05_0023E	Population over 61y	Población estimada en la tercera edad

5.2.3. Objetivos

El objetivo planteado con propósito de la validación del proceso de descubrimiento de anomalías espaciales se enuncia como sigue: Hallar condados de Estados Unidos con un comportamiento demográfico relacionado con los sectores etarios y la proporción de personas de acuerdo a su sexo biológico que presente variaciones regionales en el año 2019.

5.2.4. Implementación del proceso de explotación de Información

5.2.4.1. Preparación de los datos

Inicialmente, de la totalidad de datos obtenidos se han descartado aquellos atributos referidos a cuestiones étnicas o raciales. Posteriormente, se han descartado además datos referidos a las cantidades estimadas de población en los sectores manteniendo solo atributos referidos a las cuestiones mencionadas en el objetivo presentado en la sección previa. De esta forma restan cinco atributos que se detallan en la tabla 5.8. Cabe aclarar que el atributo FIPS ha sido generado a partir del atributo ID para realizar la integración que se detalla seguidamente.

Por otro lado, se ha eliminado el registro referido a la totalidad de Estados Unidos (ID 0100000US) así como también todos los registros asociados con regiones administrativas de Puerto Rico y de Alaska, manteniendo así una totalidad de 3220 tuplas.

En la etapa de preparación de los datos, además, se ha procedido a la limpieza

de anomalías globales. Según el objetivo planteado, las anomalías globales solo introducirían ruido a los modelos de aprendizaje no supervisado en la etapa de selección de vecindarios por lo que se ha procedido a eliminarlas utilizando el algoritmo LOF (Breunig et al., 2000) y el criterio $LOF > 2$ para considerar a las tuplas como anómalas (Kuna et al., 2009). Mediante este método se han filtrado 26 tuplas que se corresponden en su mayoría con grandes centros poblados como Washington, D.C., Denver (Colorado), Las Vegas, entre otros. En las figuras 5.8 se puede observar la distribución de valores de los atributos *Population Over 61y* y *total sex ratio* en la base de datos limpia de valores anómalos.

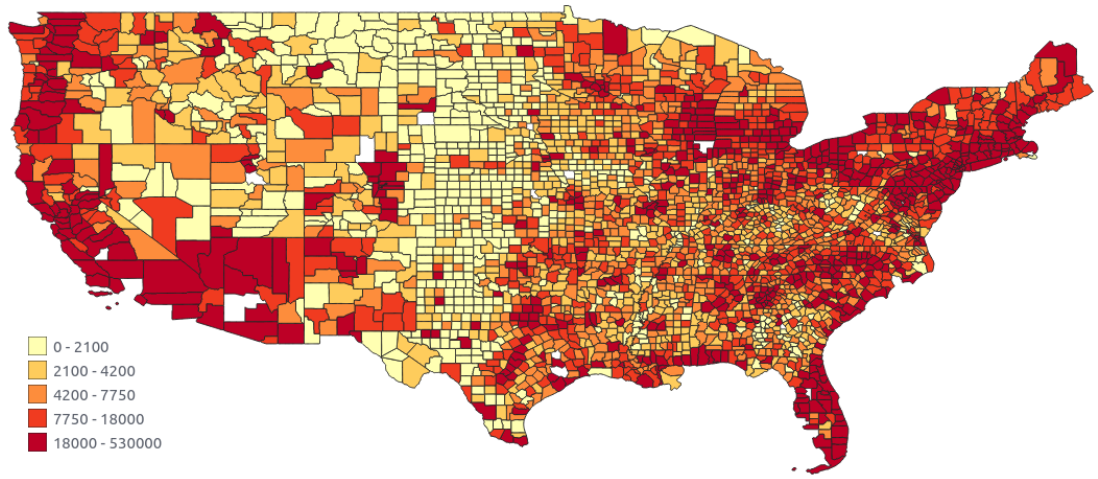
Todas estas actividades previamente mencionadas fueron ejecutadas haciendo uso del ya mencionado software para el análisis de datos Rapidminer. Los archivos resultantes en formato CSV fueron exportados para continuar con el proceso.

Para finalizar esta actividad, debido a que estos datos no se encuentran originalmente referenciados geográficamente, se han integrado haciendo uso de la herramienta QGIS 3.4 con una base de datos en formato ESRI Shapefile compuesta de polígonos para cada condado de los Estados Unidos obtenida del catálogo de datos oficiales del país (US Census Bureau, 2019). La base de datos preparada se encuentra disponible en (Rottoli, 2020d).

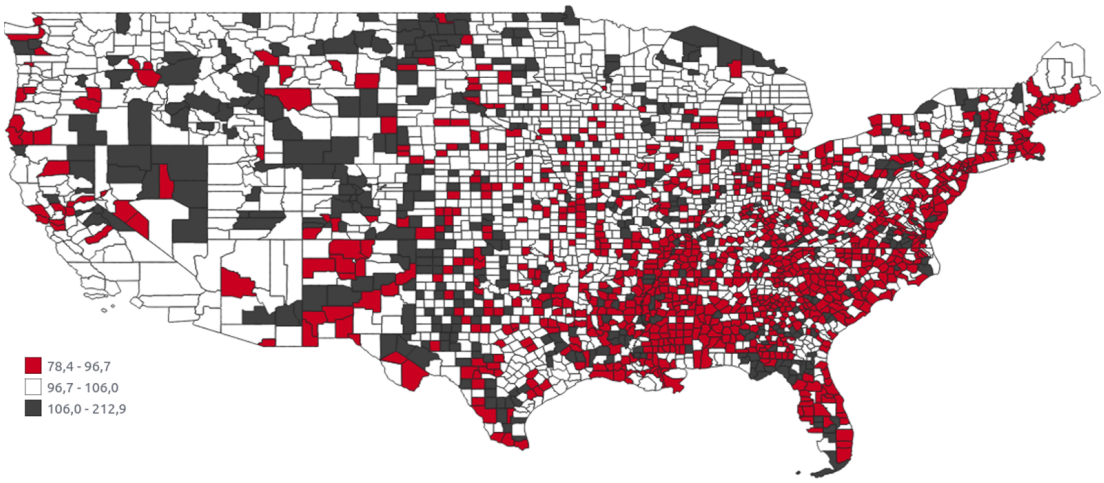
5.2.4.2. Definición de Vecindarios

La definición de vecindarios se ha realizado por similitud de los condados para disminuir la heterogeneidad espacial mediante el uso del algoritmo de regionalización SKATER (Assunção et al., 2006).

El criterio de contigüidad utilizado fue la contigüidad geográfica determinada por las fronteras naturales de los condados, esto es, dos condados son contiguos si son colindantes 5.9a. Esta condición fue utilizada para la construcción de 8 vecindarios con un mínimo de 250 objetos espaciales cada uno mediante la aplicación GeoDa con distancia euclidea y estandarización Z de los datos, obteniéndose la distribución de regiones que se observa en la figura 5.9b. Las métricas asociadas a esta regionalización se encuentran detalladas en la tabla 5.9. Un archivo CSV fue generado con los atributos utilizados y la columna adicional donde se detalla el número de región al que pertenece cada tupla.



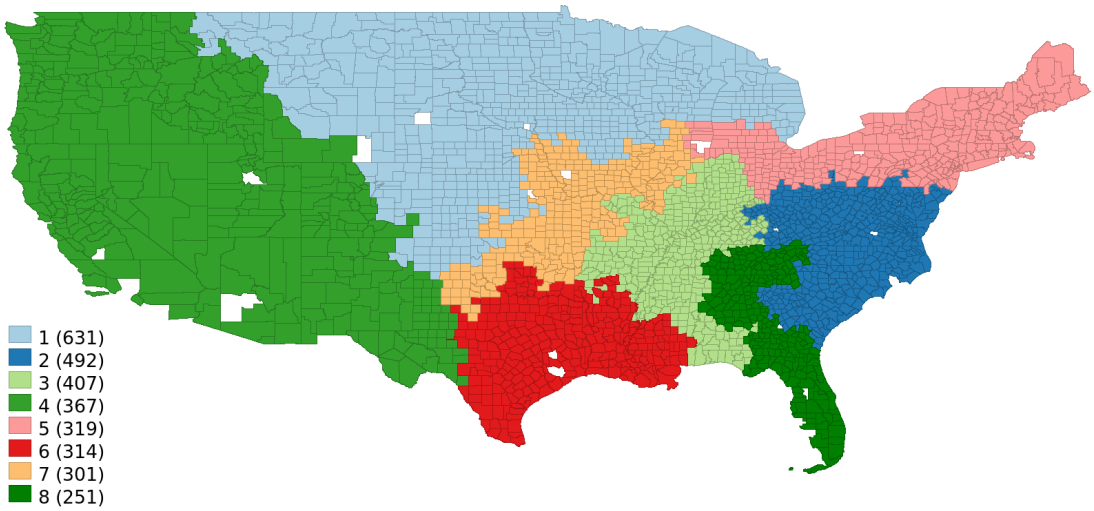
(a)



(b)

Figura 5.8: Distribución de valores de los atributos *Population Over 61y* (a) y *total sex ratio* (b) de la base de datos utilizada en la validación del proceso de detección de anomalías espaciales.

(a)



(b)

Figura 5.9: (a) Ejemplo de vecindarios de un condado de los Estados Unidos mediante el criterio de contigüidad por límites geográficos. En azul los vecinos del condado central. (b) Regiones obtenidas mediante la aplicación del algoritmo SKATER.

Tabla 5.9: Métricas de la ejecución del algoritmo de regionalización sobre los datos utilizados para la validación del proceso de descubrimiento de anomalías espaciales.

Métrica	**Valor**
Suma de cuadrados total	15405
Suma de cuadrados interna del grupo 1	1196.03
Suma de cuadrados interna del grupo 2	1226.01
Suma de cuadrados interna del grupo 3	627.899
Suma de cuadrados interna del grupo 4	3953.23
Suma de cuadrados interna del grupo 5	7243.87
Suma de cuadrados interna del grupo 6	1884.49
Suma de cuadrados interna del grupo 7	533.006
Suma de cuadrados interna del grupo 8	1982.07
Suma de cuadrados interna total	14440.5
Suma de cuadrados entre grupos	964.494
Relación entre suma de cuadrados	0.0626092

5.2.4.3. Descubrimiento de anomalías locales

El archivo CSV generado en la actividad anterior fue utilizado como entrada para el descubrimiento de anomalías locales en la aplicación Rapidminer. Para esto, primeramente, se han normalizado los atributos para evitar sesgos en la detección de anomalías. Luego, la base de datos se ha particionado según la región a la que pertenece cada dato, aplicando a cada una de estas partes del dataset original el algoritmo para la detección de anomalías Local Outlier Factor (LOF) (Breunig et al., 2000). Debido a la gran cantidad de tuplas con un valor de esta métrica mayor o igual a 2 (Kuna et al., 2009) (69 ejemplos), con fines ilustrativos en esta tesis sólo detallaremos una porción de ellas en la tabla 5.10, más precisamente aquellas con un valor mayor a 2.5.

5.2.4.4. Descripción de los vecindarios

Las tuplas anómalas fueron extraídas de la base de datos tal como detalla la actividad del proceso evaluado. Posteriormente se han descripto las regiones limpias de anomalías mediante un *script* escrito en el lenguaje de programación Python utilizando la librería Pandas, obteniendo los valores para cada atributo de los descriptores estadísticos promedio, desviación estándar, valor máximo, valor mínimo, primer, segundo y tercer cuartil. A continuación, en la tabla 5.11 se presentan los valores promedio y las desviaciones estándar de cada atributo para cada una de las regiones; el resto de los

Tabla 5.10: Outliers locales detectados en cada región con un valor de LOF mayor a 2.5. Los atributos son, en orden, Total Sex ratio, Population under 18y, Population over 17y, Population over 61.

Sex ratio	-18y	+17y	+61	Id	Región	LOF
97.5	716767	2439049	523987	0500000US06059	4	3.9
94.3	553299	2149303	501744	0500000US12086	8	3.2
89.9	608815	2026306	418541	0500000US36047	5	3.0
204.8	2471	8625	540	0500000US13053	3	2.9
94.0	21235	120138	57620	0500000US12017	8	2.8
97.2	682882	1869331	323082	0500000US48113	6	2.8
111.4	7	67	23	0500000US48301	4	2.7
103.1	9517	17893	3683	0500000US16051	4	2.7
96.0	4017	5999	919	0500000US46121	1	2.7
93.9	17641	82536	38574	0500000US12055	8	2.6
103.5	9552	19754	3650	0500000US49051	4	2.6
98.4	43138	112439	36089	0500000US49053	4	2.6
99.0	613721	1741281	385788	0500000US06065	4	2.6
94.1	473695	1865585	410891	0500000US36081	5	2.6
85.5	284	1929	1036	0500000US51091	2	2.6
102.8	7093	12796	2245	0500000US48165	4	2.6

descriptores queda disponible en el repositorio anteriormente citado para su consulta.

5.2.4.5. Descubrimiento de vecindarios anómalos

Las ocho tuplas que constituyen la descripción de las regiones o vecindarios obtenidas en la actividad del proceso de explotación de información fueron utilizadas como entrada para la detección de su grado de anormalidad en conjunto. Como ya se mencionó en el apartado correspondiente al detalle teórico del proceso en cuestión esta evaluación permite determinar si una región detectada difiere considerablemente de las otras. En una primera instancia, las métricas del algoritmo de agrupamiento (Tabla 5.9) pueden indicarnos una tendencia hacia esta anormalidad al consultar la distancia entre *clusters*, pero esta información nos resulta insuficiente para la toma de decisiones estratégicas.

En consecuencia se ha obtenido el grado de anormalidad de cada una de las regiones con base en estos descriptores haciendo uso del algoritmo *Connectivity-based Outlier Factor* o COF (Tang et al., 2002) debido a la baja densidad de puntos consi-

Tabla 5.11: Promedio de los valores de los atributos, en orden, Total Sex ratio, Population under 18y, Population over 17y, Population over 61, para cada una de las regiones.

	Sex ratio	**-18y**	**+17y**	**+61y**
1	102.5 ± 9.0	9641.7 ± 29539.0	32283.2 ± 98176.2	7969.0 ± 22205.4
2	98.0 ± 9.6	19178.0 ± 36665.8	65302.1 ± 118657.8	15370.7 ± 24020.0
3	99.1 ± 9.5	10411.5 ± 21233.1	34649.1 ± 66630.8	8558.9 ± 14444.3
4	104.8 ± 13.6	25274.6 ± 53752.8	80645.9 ± 167348.1	19381.3 ± 36875.2
5	98.2 ± 5.3	48418.7 ± 65689.2	172728.7 ± 228572.8	42813.3 ± 53396.7
6	100.6 ± 14.0	17981.7 ± 36948.8	54188.2 ± 104808.6	12134.3 ± 19508.7
7	101.0 ± 10.0	12654.1 ± 26491.1	40909.5 ± 84748.0	9955.0 ± 19675.0
8	100.5 ± 14.4	21043.9 ± 41445.4	75347.4 ± 150240.8	19918.1 ± 39979.4

derados. Para esto se hizo uso de Rapidminer, tal como se ha procedido en anteriores oportunidades, con los parámetros k=2 y distancia euclidiana mixta. Como resultado la región número 5 es clasificada como anómala por poseer un valor de COF igual a 3.125 en contraste con las demás regiones que poseen valores por debajo de 1.04.

5.2.5. Evaluación de los resultados

A partir de la tabla 5.11 que posee la descripción de las ocho divisiones que se han realizado sobre los condados de Estados Unidos presentes en el dataset en cuestión, y de la figura 5.11, que muestra estos valores enfrentados en una matriz de gráficos de dispersión, se puede evidenciar una tendencia lineal entre la cantidad de población de los tres sectores etarios. Aparentemente una mayor cantidad promedio de población se correlaciona con una mayor cantidad promedio de personas en los fragmentos de la mayoría de edad y la tercera edad. No existe una correlación aparente entre el atributo Total Sex Ratio y los demás.

En relación a lo expuesto, los *clusters* toman distintos valores en esta línea de regresión, variando considerablemente la desviación estándar en cada caso. Esto se puede observar en la figura 5.10, que muestra para cada región un gráficos de bigotes (*boxplot*) en el que se puede ver la distribución de valores individualmente para cada atributo. Además, en rojo se aprecian los *outliers* detectados mediante el método expuesto con anterioridad. Cabe destacar que no fueron detallados los valores que exceden los límites superior e inferior de cada diagrama, esto es, los *outliers* individuales de cada atributo para cada grupo. En las tuplas anómalas detectadas se tiene en cuenta

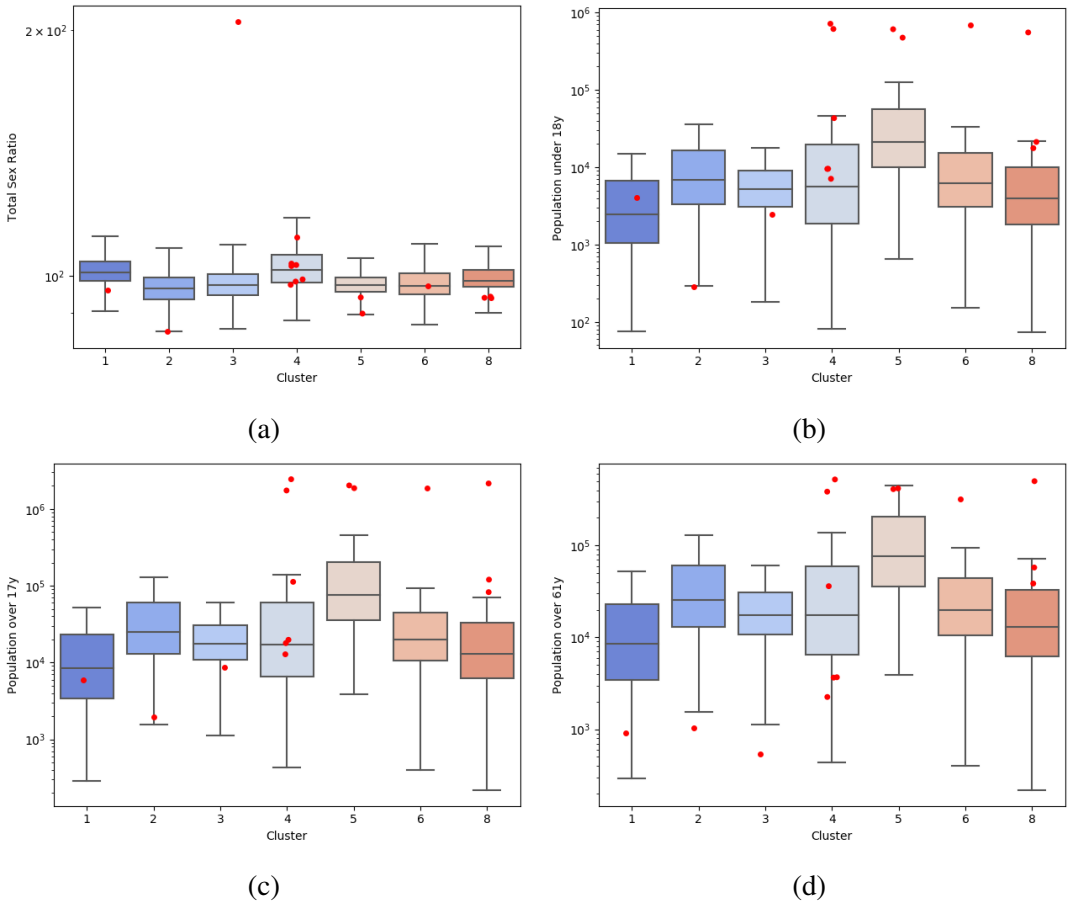


Figura 5.10: Gráficos de bigotes con la distribución de valores de los atributos (a) *Total Sex Ratio*, (b) *Population under 18y*, (c) Population over 17y y (d) *Population over 61y* en escala logarítmica para cada uno de las regiones o *clusters* generados. En rojo los datos considerados anómalos. La región 7 no se muestra por no poseer valores anómalos.

la relación existente entre los valores de cada una de estas dimensiones.

En cuanto a la región 1, se ha detectado una única anomalía (ID 0500000US46121) utilizando la condición especificada con anterioridad. Esta anomalía se encuentra a menos de una desviación estándar de cada uno de los parámetros, sin embargo se puede notar como, a pesar de poseer una tamaño de población menor de edad y población mayor de edad dentro del rango intercuartil, posee una población de la tercera edad escasa, por debajo del primer cuartil. Los valores exactos se pueden observar en la tabla 5.10, presentada anteriormente. La correlación entre las tres variables que se ha mencionado anteriormente no se verifica en este caso.

En la región 2 también se ha determinado la presencia de un único valor anómalo

(ID 0500000US51091). En este caso sus valores para cada uno de los atributos se encuentran fuera o muy cercanos a los límites inferiores de cada gráfico de bigotes. Esto implica un tamaño de la población reducido en comparación al resto de condados de la región y una predominancia de personas del sexo femenino mayor a la esperada (85 personas masculinas por cada 100 femeninas). Es importante destacar la importancia de la autocorrelación espacial en este caso particular: los valores que vuelven anómalos a esta tupla resultan normales en otras regiones, tal como las regiones 1 y 4.

La región 3 posee al igual que las anteriores un solo condado de naturaleza anómala (ID 0500000US13053), en este caso relacionado casi exclusivamente a dos factores: su considerablemente alta cantidad de hombres por cada 100 mujeres (204.8) y su escasa población de la tercera edad, aunque su total de habitantes también es reducido. Se sugiere indagar en la posibilidad de que el valor del primer atributo indicado sea un error.

Por otro lado, en contraste con lo anteriormente detallado, la región 4 posee 7 condados con un comportamiento que desvía del normal. En primera instancia, este comportamiento anómalo no se debe al atributo Total Sex Ratio, como se observa en el gráfico de bigotes correspondiente a esta región. Dos de estas tuplas poseen una población total muy por encima de lo esperado para este conjunto de condados (ID 0500000US48113, 0500000US06065), con una población mayor de edad superior al millón y medio de habitantes, excediendo al valor promedio del *cluster* en aproximadamente 10 desviaciones estándar, fenómeno que se observa además en los atributos restantes. Por otro lado, las anomalías correspondientes a los condados ID 0500000US16051, 0500000US4905 y 0500000US48165 presentan una población de adultos mayores por debajo de lo esperado para la población mayor de edad y menor de edad que poseen. Este fenómeno también se verifica para el condado ID 0500000US06059, aunque la población de cada segmento etario sea normal, el comportamiento conjunto hace que esta subdivisión administrativa sea anómala. Por último, la tupla cuyo ID es 0500000US48301 posee valores excepcionalmente bajos. Cabe la posibilidad de que estos valores sean erróneos.

Por último, las anomalías presentes en las regiones 5, 6 y 8 muestran poseer una cantidad de población mayor al normal, principalmente en la región 5, mientras que este comportamiento en las regiones 6 y 8 es más pronunciado en el segmento etario de los mayores de 18 años. Se puede ver además que los condados anómalos de los sectores 5 y 8 poseen una ligera reducción en la cantidad de personas del sexo masculino

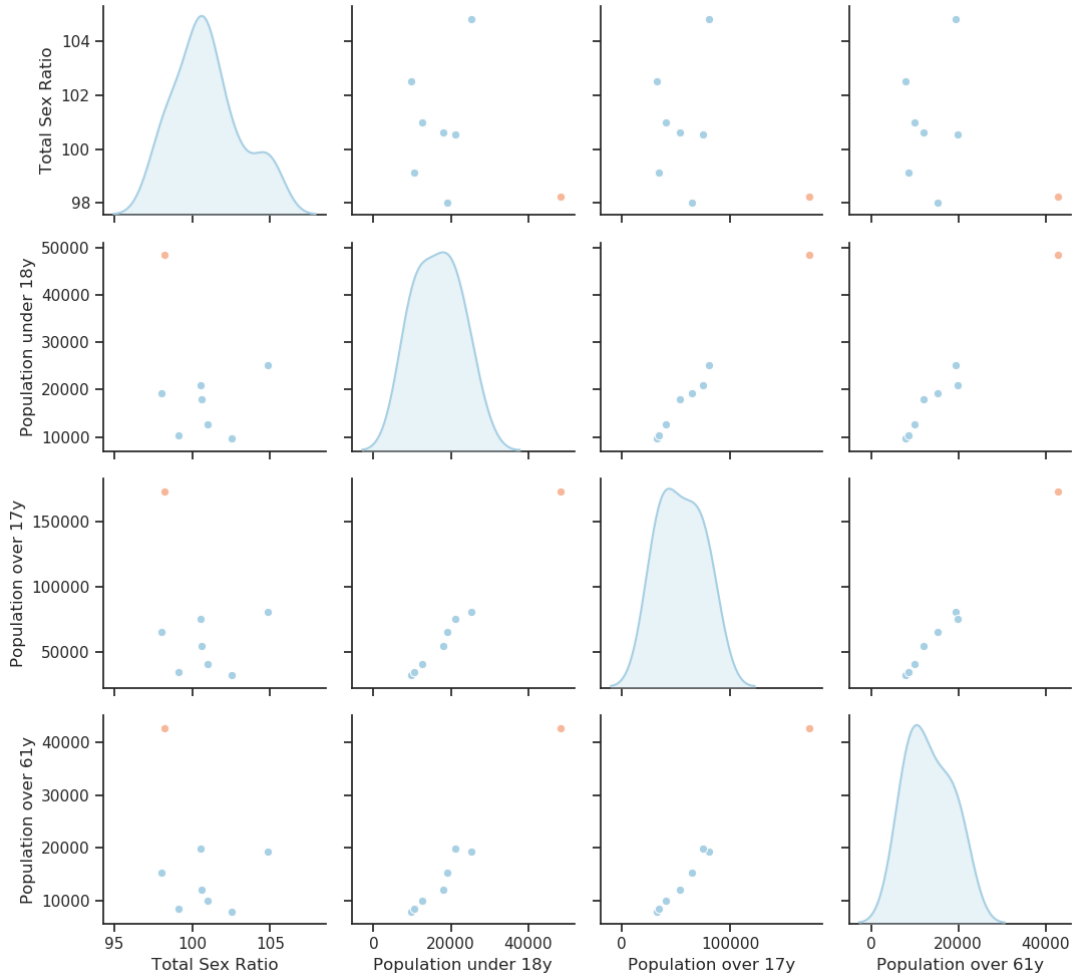


Figura 5.11: Matriz de dispersión de las regiones en función de los valores promedio de sus atributos. En azul las regiones normales y en naranja la región anómala detectada.

que lo que se espera como normal.

Por otro lado, la región 5, que ha sido clasificada como anómala, posee condados con un promedio de población más alto que el de las demás regiones, cuestión que se puede observar tanto en los gráficos de bigotes de la figura 5.10 y en los diagramas de dispersión de la figura 5.11. Esta región se corresponde en una gran parte con la costa este de los Estados Unidos y sus centros más poblados.

5.3. Descubrimiento de asociaciones espaciales

A continuación se describe el contexto del problema utilizado en la validación del proceso de descubrimiento de asociaciones espaciales (Sección 5.3.1), se detallan los datos utilizados para tal propósito (Sección 5.3.2), se plantea el objetivo de minería de datos asociado al problema (Sección 5.3.3), se desarrolla la implementación del proceso en cuestión (Sección 5.3.4) y por último se evalúan los resultados haciendo uso de la opinión de un experto (Sección 5.3.5).

5.3.1. Contexto del problema

La ciudad de Concepción del Uruguay se ubica en la costa occidental del río Uruguay, en la pronvincia de Entre Ríos, en Argentina. Esta urbe de 72528 habitantes según el censo nacional de población del año 2010, es la cabecera del departamento homónimo y es conocida nacional e internacionalmente como ”La Histórica”, debido a su influencia social y política desde principios de 1800, además de haber sido la sede de estudio de múltiples personajes de renombre de la historia argentina, como los ex presidentes Julio Argentino Roca, Victorino de la Plaza y Arturo Frondizi, del presidente paraguayo Benigno Ferreira, gobernadores y ministros de múltiples provincias y personajes destacados de la cultura nacional, como la Dra. Teresa Ratto, segunda mujer médica del país. Además, Concepción del Uruguay se destaca por haber sido el lugar de promulgación de la constitución nacional del año 1853 en manos del Gral. Justo José de Urquiza (Urquiza Almandor, 2002a,b,c).

La validación del proceso de explotación de información para el descubrimiento de asociaciones espaciales se realiza sobre una base de datos espacialmente referenciada en el sector céntrico de la ciudad con propósitos de determinar relaciones espaciales entre distintos tipos de servicios para gestión urbanística.

5.3.2. Datos utilizados

Los datos utilizados en la validación del proceso de explotación de información para el descubrimiento de asociaciones espaciales fueron obtenidos a través de múltiples consultas sobre la aplicación Overpass Turbo (Raifer, 2020), una herramienta de filtrado de datos basado en la web para OpenStreetMap (`https://www.openstreetmap.org`).

Las consultas realizadas permitieron obtener información sobre los puntos espaciales relacionados con distintos servicios y ubicaciones particulares de la ciudad en su sector céntrico. Por cada punto, además de su ubicación, se tiene un atributo no espacial categórico denominado "amenity" en el que consta el tipo de objeto espacial, servicio o ubicación de interés representado. Los posibles valores que pueden tomar este atributo y la cantidad de instancias de cada uno de ellos se encuentran detallados en la tabla 5.12.

Asimismo, se han agregado dos capas de datos de referencia: una capa de objetos espaciales lineales por cada calle principal de la ciudad que pudiera ser relevante para el estudio, y una capa de polígonos ubicados sobre los espacios verdes y de esparcimiento del sector céntrico. Cabe aclarar que solo se consideraron aquellos de interés para el estudio y se ignoraron aquellos que han resultado irrelevantes para el mismo según opinión del experto. Asimismo, tanto las líneas como los polígonos han sido categorizados al igual que los puntos. Estas categorías se detallan en la tabla 5.13 La distribución de tanto los puntos como de las líneas y los polígonos se puede apreciar en la figura 5.12.

5.3.3. Objetivos

El objetivo de minería de datos planteado para la validación del proceso de descubrimiento de asociaciones espaciales consiste en la determinación de asociaciones entre las distintos servicios del sector céntrico de Concepción del Uruguay, sus calles principales y los espacios verdes públicos. Son de especial interés la cercanía entre estos objetos y su orientación relativa en el espacio.

5.3.4. Implementación del proceso de explotación de Información

5.3.4.1. Preparación de los datos

Los datos puntuales utilizados fueron obtenidos mediante consultas realizadas sobre la base de datos de OpenStreetMaps, tal como se indicó en la sección 5.3.2, creando una única capa SHP mediante el uso de la aplicación QGIS, al igual que en las validaciones anteriores.

Posteriormente se crearon de forma manual tanto las líneas como los polígonos de

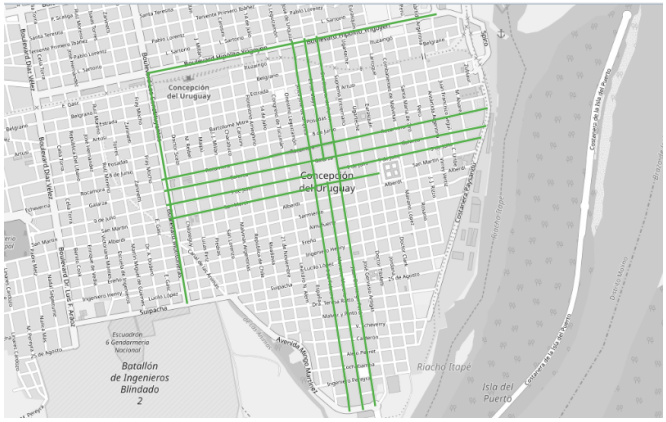
Clase	Instancias	Clase	Instancias
Almacén	18	Hotel	17
Banco	9	Iglesia	7
Bar	10	Joyería	3
Barbería	6	Lavandería	3
Biblioteca	1	Librería	11
Bicicletería	4	Mercado	30
Bomberos	1	Monumento	7
Carnicería	9	Panadería	15
Casino	1	Policía	2
Cine	1	Productos de Limpieza	3
Clínicas	17	Pub	3
Correo	3	Restaurante	30
Disquería	1	Supermercado	11
Escuela	15	Taxi	2
Estación de Bus	1	Tienda de Deportes	4
Farmacia	13	Tienda de Mascotas	3
Ferretería	6	Tienda de Ropa	32
Florería	1	Tienda de Tecnología	14
Fotografo	3	Tienda General	2
Geriátrico	4	Turismo	3
Gimnasio	11	Zapatería	7
Heladería	8		

Tabla 5.12: Clases de puntos espaciales y cantidad de instancias de cada clase utilizadas en la validación del proceso de explotación de información para el descubrimiento de asociaciones espaciales.

Tipo	Clase	Cantidad
Línea	Boulevard	3
Línea	Peatonal	1
Línea	Calle Principal	11
Polígono	Plaza	5
Polígono	Predio	1

Tabla 5.13: Clases de líneas y polígonos espaciales y cantidad de instancias de cada clase utilizadas en la validación del proceso de explotación de información para el descubrimiento de asociaciones espaciales.

(a)



(b)

(c)

Figura 5.12: Distribución de objetos espaciales de la base de datos utilizada para la validación del proceso de descubrimiento de asociaciones espaciales. (a) Puntos; (b) líneas; (c) polígonos.

interés. De esta forma se obtienen tres archivos de ESRI Shapefile, uno por cada tipo de dato espacial que son utilizados como entrada de la próxima actividad del proceso. La base de datos original y el repositorio preparado utilizado en esta prueba pueden encontrarse en (Rottoli, 2020b).

5.3.4.2. Definición de vecindarios

En esta validación resulta relevante que los patrones descubiertos apliquen a toda la zona considerada, por lo que se considera todo el espacio de datos como un único vecindario.

En este sentido la definición de vecindarios nos permite segmentar el espacio de búsqueda cuando es relevante pero no nos limita obligándonos a hacerlo en cada oportunidad, por lo que tanto este proceso como los anteriores mencionados son adaptables a los distintos problemas de negocio.

5.3.4.3. Modelado de relaciones espaciales

Como se especifica en el objetivo de minería de datos descripto anteriormente, son de especial interés las relaciones euclidianas de vecindad y direccionales relacionadas con cuatro puntos cardinales principales. La implementación de las mismas fue realizada mediante el uso de teoría de conjuntos difusos tal como se menciona en la sección 4.1.2.

En primer lugar, un par ordenado de puntos espaciales pertenece a la relación de vecindad con un grado de pertenencia igual a 1 si su distancia es menor a 50 metros y en grado de pertenencia igual a $\frac{50-x}{30} + 1$ si la distancia x se encuentra entre 50 y 80 metros. A una distancia mayor los dos puntos no se consideran vecinos (Fig. 5.13a).

De forma similar, un punto espacial es vecino de una línea en el espacio con grado de pertenencia igual a 1 si su distancia es menor a 30 metros e igual a $\frac{30-x}{50} + 1$ si la distancia x se encuentra entre 30 y 80 metros. De igual forma, a una distancia mayor el grado de pertenencia es cero. Cabe aclarar además que la distancia se calcula entre el punto y el punto más cercano de la línea evaluada (Fig. 5.13b).

Por otro lado, el grado de pertenencia de los pares ordenados (A, B) a las relaciones direccionales binarias definidas por los predicados $Norte$,Sur, $Este$ y $Oeste$, donde A y B son puntos espaciales, se calcularon siempre que la distancia euclídea entre ellos

no supere los 500 metros utilizando las fórmulas que se detallan en las ecuaciones 5.2, 5.3, 5.4 y 5.5 (Fig. 5.14). Asimismo, las relaciones entre puntos y polígonos fueron calculadas de la misma manera pero considerando el centroide de este último como punto de referencia.

De esta forma, mediante un programa realizado mediante el lenguaje de programación Python se calcularon los grados de pertenencia de los pares de objetos a las relaciones en cuestión utilizando como entrada las capas SHAPEFILE creadas en el paso previo y generando un archivo JSON de salida para ser utilizado en la actividad que procede.

$$\mu_{N\tilde{o}rte}(A,B) = \begin{cases} 1 & \text{si } A_x = B_x \wedge A_y > B_y \\ |\frac{2}{\pi}\tan^{-1}(\frac{A_y - B_y}{A_x - B_x})| & \text{si } A_x \neq B_x \wedge A_y > B_y \\ 0 & \text{en caso contrario} \end{cases} \tag{5.2}$$

$$\mu_{S\tilde{u}r}(A,B) = \begin{cases} 1 & \text{si } A_x = B_x \wedge A_y < B_y \\ |\frac{2}{\pi}\tan^{-1}(\frac{A_y - B_y}{A_x - B_x})| & \text{si } A_x \neq B_x \wedge A_y < B_y \\ 0 & \text{en caso contrario} \end{cases} \tag{5.3}$$

$$\mu_{E\tilde{s}te}(A,B) = \begin{cases} 1 & \text{si } A_y = B_y \wedge A_x > B_x \\ |\frac{2}{\pi}\tan^{-1}(\frac{A_x - B_x}{A_y - B_y})| & \text{si } A_y \neq B_y \wedge A_x > B_x \\ 0 & \text{en caso contrario} \end{cases} \tag{5.4}$$

$$\mu_{O\tilde{e}ste}(A,B) = \begin{cases} 1 & \text{si } A_y = B_y \wedge A_x < B_x \\ |\frac{2}{\pi}\tan^{-1}(\frac{A_x - B_x}{A_y - B_y})| & \text{si } A_y \neq B_y \wedge A_x < B_x \\ 0 & \text{en caso contrario} \end{cases} \tag{5.5}$$

5.3.4.4. Minado de subgrafos frecuentes

Una vez obtenido el modelado difuso de las relaciones espaciales de interés para la inteligencia de negocio se procedió a la extracción de subgrafos frecuentes haciendo uso del algoritmo Subdue (Holder et al., 1994) implementado en la herramienta *Subdue Graph Miner* (Holder, 2020).

Los siete patrones más relevantes haciendo uso de la métrica de compresión de grafos son listados en la tabla 5.14. Asimismo, los grados de pertenencia a las relaciones que están involucradas en cada patrón se obtuvieron utilizando la T-Norma Producto

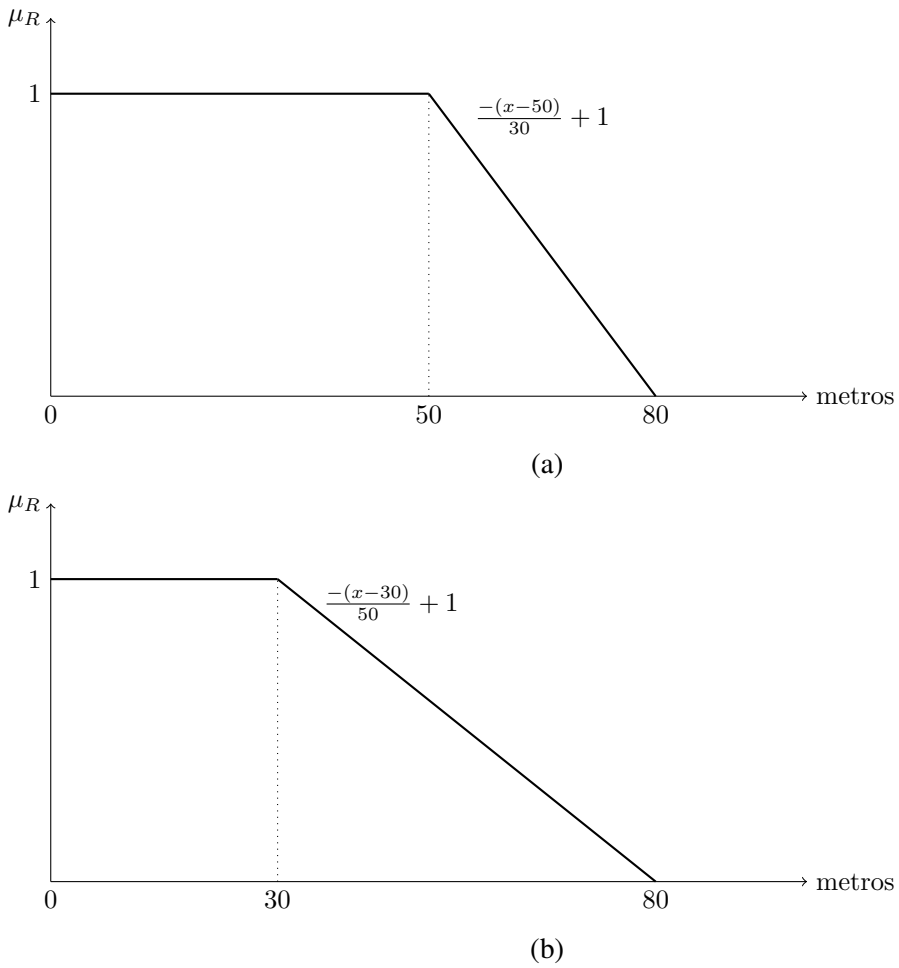


Figura 5.13: Función de pertenencia de la relación de vecindad entre (a) dos puntos espaciales y (b) un punto espacial y una línea espacial.

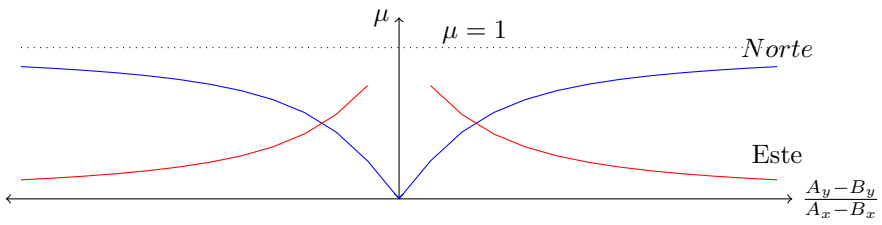


Figura 5.14: Función de pertenencia de las relaciones definidas por los predicados Norte(A,B) y Este(A,B) en función de la pendiente entre ambos puntos. Cabe aclarar que la relación Sur es equivalente a la relación inversa de la relación Norte menos la relación identidad al igual que la relación Oeste es la relación inversa de la relación Este menos la relación identidad.

#	Patrón	Comp.	Inst.
1	$Tienda\,de\,Ropa(X_1) \wedge CallePrincipal(X_2) \wedge \tilde{Vecino}(X_1, X_2)$	0.023	44
2	$Tienda\,de\,Ropa(X_2) \wedge Tienda\,de\,Ropa(X_2) \wedge \tilde{Vecino}(X_1, X_2)$	0.023	44
3	$Restaurante(X_1) \wedge Calle\,Principal(X_2) \wedge \tilde{Vecino}(X_1, X_2)$	0.021	41
4	$Tienda\,de\,Ropa(X_1) \wedge Zapateria(X_2) \wedge \tilde{Vecino}(X_1, X_2)$	0.013	27
5	$Escuela(X_1) \wedge Restaurante(X_2) \wedge Panaderia(X_3) \wedge Farmacia(X_4) \wedge Almacen(X_5) \wedge Calle\,Principal(Y) \wedge \tilde{Vecino}(X_1, Y) \wedge \tilde{Vecino}(X_2, Y) \wedge \tilde{Vecino}(X_3, Y) \wedge \tilde{Vecino}(X_4, Y) \wedge \tilde{Vecino}(X_5, Y)$	0.013	6
6	$Escuela(X_1) \wedge Restaurante(X_2) \wedge Panaderia(X_3) \wedge Farmacia(X_4) \wedge Mercado(X_5) \wedge Calle\,Principal(Y) \wedge \tilde{Vecino}(X_1, Y) \wedge \tilde{Vecino}(X_2, Y) \wedge \tilde{Vecino}(X_3, Y) \wedge \tilde{Vecino}(X_4, Y) \wedge \tilde{Vecino}(X_5, Y)$	0.013	6
7	$Escuela(X_1) \wedge Restaurante(X_2) \wedge Panaderia(X_3) \wedge Farmacia(X_4) \wedge Calle\,Principal(Y) \wedge \tilde{Vecino}(X_1, Y) \wedge \tilde{Vecino}(X_2, Y) \wedge \tilde{Vecino}(X_3, Y) \wedge \tilde{Vecino}(X_4, Y)$	0.012	7

Tabla 5.14: Patrones de asociación descubiertos utilizando el proceso de descubrimiento de asociaciones espaciales.

de Hamacher (Eq. 4.5) por cada instancia. En la tabla 5.15 se presentan como una segunda métrica de interés el soporte de cada tipo de objeto espacial que interviene en cada patrón. Por último, en la figura 5.15 se muestra la distribución de los grados de pertenencia calculados para cada asociación espacial.

5.3.5. Evaluación de los resultados

En primer lugar se destaca la prevalencia de asociaciones que incluyen la relación de vecindad entre puntos y entre puntos y líneas espaciales. Ningún patrón incluye ninguna de las relaciones direccionales que se han considerado en el trabajo. Se deberá indagar en futuras iteraciones la existencia o ausencia de estas asociaciones en la base de datos utilizando estrategias acordes a tal fin, como incluir una base de datos más completa y nueva clase de objetos espaciales de referencia.

Dicho lo anterior y considerando la información obtenida y mostrada en las tablas 5.14 y 5.15 y en la figura 5.15, 44 subgrafos del modelo gráfico de relaciones espaciales

Patrón	Clase	Instancias únicas	Soporte
1	Tienda de Ropa	29	0.902
	Calle Principal	8	0.727
2	Tienda de Ropa	29	0.625
3	Restaurante	25	0.833
	Calle Principal	11	1.000
4	Tienda de Ropa	12	0.375
	Zapatería	6	0.857
5	Escuela	6	0.400
	Restaurante	6	0.200
	Farmacia	6	0.461
	Panadería	6	0.400
	Almacén	6	0.300
	Calle Principal	6	0.545
6	Escuela	6	0.400
	Restaurante	6	0.200
	Farmacia	6	0.461
	Panadería	6	0.400
	Mercado	6	0.200
	Calle Principal	6	0.545
7	Escuela	7	0.466
	Restaurante	7	0.233
	Farmacia	7	0.538
	Panadería	7	0.466
	Calle Principal	7	0.636

Tabla 5.15: Soporte de las clases de objetos espaciales en cada patrón de asociación descubierto

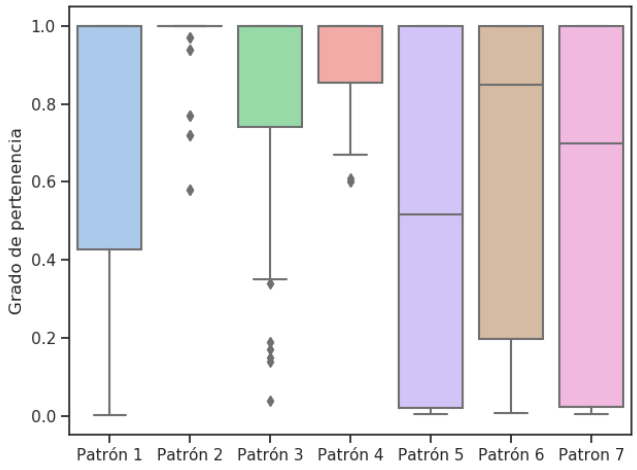


Figura 5.15: Distribución de los grados de pertenencia de las instancias que cumple con cada patrón de asociación.

poseen el mismo patrón: un 90.2 % de las tiendas de ropa (29 instancias) se encuentran vecinas a una calle principal, siendo el porcentaje de calles principales involucradas del 72.7 % (8 instancias). La mediana del grado de pertenencia a la relación de vecindad es igual a 1, por lo que más del 50 % de las instancias de la relación se encuentran a menos de 30 metros de separación.

El segundo patrón también resume 44 subgrafos e involucra a dos instancias de puntos espaciales de tiendas de ropa. En este caso el 62.5 % de las tiendas de ropa son vecinas de otra tienda de ropa, con una distribución de grados de pertenencia de las instancias del patrón a la relación de vecindad muy poco dispersa, siendo el valor promedio igual a 0.953, la mediana igual a 1 y la desviación estándar igual a 0,11. En este caso, un 62 % de las tiendas de ropa se pueden encontrar casi con certeza a unos 50 metros de distancia de otra tienda de ropa. Los valores anómalos que se observan en el gráfico de bigotes se encuentran por encima del valor 0.5.

En cuanto al tercer patrón de asociación descubierto, se nota cierta similitud con el primero: en todas las calles principales podemos encontrar un restaurante en un radio menor a 80 metros de la misma, estando involucrados el 80.3 % de los restaurantes de la base de datos. En este caso, la distribución de los grados de pertenencia también posee una mediana igual a 1 pero el valor promedio desciende hacia 0.81, siendo la desviación estándar igual a 0.3. Además, el primer cuartil se corresponde con el valor 0.74, por lo que el 75 % de los restaurantes involucrados se encuentran a menos de 58 metros de distancia de una calle principal.

El cuarto patrón involucra a las zapaterías y a las tiendas de ropa en un vecindario de 80 metros unas de otras, con un valor promedio de pertenencia a la relación de vecindad igual a 0.91, con una desviación estándar igual a 0.13 y mediana 1. Este patrón de asociación aplica para una escasa cantidad de tiendas de ropa, 37.5 %, pero para una gran cantidad de las zapaterías consideradas (85.7 %). Esto implica que es probable encontrar una tienda de ropa en un vecindario acotado de las zapaterías mas encontrar una zapatería en los vecindarios de las tiendas de ropa es menos probable.

Los últimos 3 predicados compuestos aplican con un bajo soporte para cada una de las instancias de las clases de objetos espaciales que intervienen, pero representan una tasa de compresión elevada, por lo que es un patrón recurrente.

En tanto el primer como en el segundo caso se interpreta que alrededor del 54.5 % de las calles principales es posible encontrar en un radio de 80 metros las siguientes clases de servicios: escuela, restaurante, farmacia, panadería y almacén o mercado. El grado de pertenencia a la relación difusa de vecindad, calculada con la T-norma previamente mencionada posee una gran dispersión, siendo el promedio, la mediana y la desviación estándar respectivamente igual a 0.5, 0.51 y 0.53 en el primer caso y 0.62, 0.85 y 0.48 en el segundo caso.

Por último, en cuanto al séptimo patrón, agrega una última instancia de subgrafo frecuente al ignorar los almacenes y los mercados, por lo que se describe que alrededor del 63.6 % de las calles principales podemos encontrar una escuela, un restaurante, una farmacia y una panadería, siendo los grados de pertenencia menores al sexto patrón, pero mayores al primero: 0.53 el promedio, 0.7 la mediana y 0.49 la desviación estándar.

5.4. Descubrimiento de patrones de co-localización

En esta sección del capítulo 5 se presenta la validación del proceso de descubrimiento de patrones de co-localización. Para esto se presenta el problema abordado (Sección 5.4.1), los datos utilizados para las pruebas (Sección 5.4.2), se describen los objetivos de minería de datos a satisfacer (Sección 5.4.3), se describe la forma en las que las distintas actividades del proceso fueron implementadas (Sección 5.4.4) y por último se evalúan los resultados obtenidos haciendo uso de un experto en el área (Sección 5.4.7).

5.4.1. Contexto del problema

La Ciudad y Condado de San Francisco (*City and County of San Francisco*), es la cuarta ciudad más poblada de California y la treceava de Estados Unidos, con aproximadamente 900000 habitantes en 2013. Cuenta además con la segunda densidad de población más alta del país (U.S. Census Bureau, 2020a).

El turismo es la principal actividad económica de la región, por lo que la seguridad es una de los factores más cuidados de la ciudad. Para esto, el departamento de policía de la ciudad provee estadísticas actualizadas de su situación criminológica, así como también datos abiertos para análisis varios.

El proceso de explotación de información es demostrado utilizando estas bases de datos para la búsqueda de patrones de co-localización en zonas de alta densidad de ocurrencia de crímenes.

5.4.2. Datos utilizados

Los datos utilizados para la validación del proceso de descubrimiento de patrones de co-localización fueron obtenidos de (Center for Spatial Data Science, 2017) y consisten en incidentes ocurridos desde el día 1 de julio hasta el día 31 de diciembre del año 2012 en la ciudad de San Francisco, Estados Unidos. Los datos provienen del sistema de reporte de crímenes del departamento de policía de San Francisco y fueron recolectados el día 10 de enero de 2013, preparados por el *Center for Spatial Data Science* y actualizados en el año 2017.

Los datos en cuestión fueron puestos públicos bajo licencia Creative Commons li-

Tabla 5.16: Atributos en común entre los cuatro archivos de datos SHAPEFILE sobre los incidentes ocurridos en la ciudad de San Francisco.

Atributo	Tipo	Descripción
IncidntNum	Numérico	Identificador único de cada reporte
X_pr	Numérico	Longitud proyectada de la ubicación del crimen
Y_pr	Numérico	Latitud proyectada de la ubicación del crimen
Category	Nominal	Tipo de crimen, uno por cada archivo de datos
Descript	Nominal	Descripción del crimen en categorías preestablecidas
DayOfWeek	Nominal	Día de la semana del reporte del crimen
PdDistrict	Nominal	Distrito de policía
Resolution	Nominal	Resultado del reporte
Location	Texto	Dirección donde el crimen ocurrió
Date	Fecha	Fecha del reporte
Time	Hora	Hora del reporte
X	Numérico	Longitud de la ubicación del crimen
Y	Numérico	Latitud de la ubicación del crimen

cense (CC0 1.0 Universal) en cuatro archivos SHAPEFILE en el portal de datos abiertos de la ciudad mencionada. Estos archivos de datos contienen datos puntuales de robo de vehículos (*Car thefts*, 607 observaciones), drogas (*Drugs*, 3897 observaciones), robos (*Robberies*, 2761 observaciones) y vandalismo (*Vandalism*, 3430 observaciones). Cada uno de estos cuatro archivos cuenta como mínimo con los atributos que se muestran en la tabla 5.16.

5.4.3. Objetivos

Se determina como objetivo de minería de datos del proyecto de explotación de información la determinación de hechos delictivos que generalmente ocurran próximos en el espacio en el año 2012. Debido a la amplia cantidad de hechos delictivos en toda la ciudad, interesan las zonas donde la concentración de estos hechos resulta elevada.

Interesa para tal propósito que dos hechos delictivos no se encuentren a una distancia superior a los 100 metros para ser considerados como vecinos. Se ignorará el factor temporal para los propósitos de esta demostración.

5.4.4. Implementación del proceso de explotación de Información

5.4.4.1. Preparación de los datos

En la etapa de preparación de los datos los cuatro archivos de datos obtenidos del repositorio previamente mencionado fueron integrados haciendo uso de la herramienta QGIS (Qgis, 2020), ya utilizada en las secciones anteriores. De esta forma, una única capa de datos puntuales quedó conformada con 13472 puntos descriptos por los atributos mencionados en la tabla 5.16. En la figura 5.16 puede observarse la zona en la que los datos fueron obtenidos y la distribución de los mismos.

Ediciones adicionales no fueron necesarias debido a que las bases de datos se encuentran limpias, sin anomalías y sin datos extraviados. La base de datos preparada puede ser accedida en (Rottoli, 2020c).

5.4.4.2. Definición de vecindarios

En las implementaciones de los procesos de explotación de información anteriores, distintas estrategias fueron utilizadas para la generación de vecindarios según el problema de negocio abordado. En esta oportunidad, debido a que el objetivo planteado especifica la necesidad de evaluar las zonas de altas concentraciones de crímenes, se ha optado por la utilización de búsqueda de zonas calientes mediante algoritmos basados en métricas estadísticas.

Para este trabajo se ha seleccionado el algoritmo G_i^* de Getis-Ord (Getis and Ord, 1992; Ord and Getis, 1995) disponible en la extensión "Hotspot Analysis" para QGIS (Oxoli et al., 2019) que considera la autocorrelación espacial de los datos.

Para esto, por cada dato se calcula el valor G_i^*, el cual es un valor z, según la definición 5.6, donde x_j es el valor del atributo j, $w_{i,j}$ es el valor del ponderador espacial entre los puntos i y j, n es el número total de puntos, y $\bar{X}$ y S se calculan según las ecuaciones 5.7 y 5.8 respectivamente.

Como resultado de este proceso, dos zonas con alta densidad de puntos fue hallada y pueden observarse en la figura 5.17. Las regiones constan de puntos donde $1.96 \leq z < 2.58$ y $0.01 < p - value \leq 0.05$ (puntos con un 95 % de confianza) y los puntos donde $z \geq 2.58$ y $p - value \leq 0.01$ (puntos con un 99 % de confianza).

De esta forma fue generado un nuevo archivo con 1119 puntos utilizado para el

(a)

(b)

Figura 5.16: (a) Región de San Francisco en donde se desarrolla el estudio. (b) Distribución de puntos en la capa resultante de integrar las fuentes de datos.

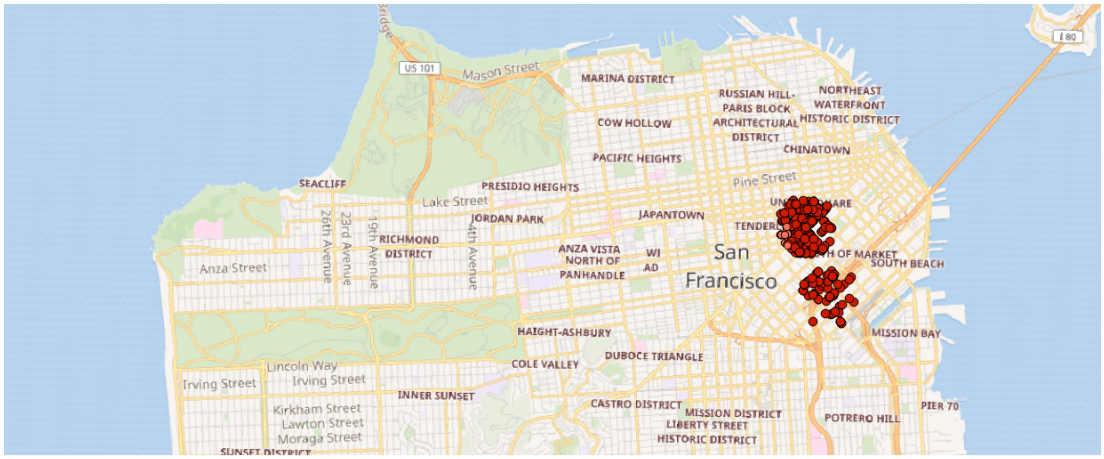


Figura 5.17: Zonas calientes según la concentración de puntos dada por la métrica G_i* de los crímenes en la ciudad de San Francisco. Se notan dos zonas calientes muy cercanas entre sí en la zona noreste de la ciudad.

cálculo de las relaciones de vecindad y la generación de transacciones que se encuentra disponible en el repositorio adjunto (Rottoli, 2020c).

$$G_i^* = \frac{\sum_{i=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2\right]}{n-1}}} \tag{5.6}$$

$$\bar{X} = \frac{\sum_{j=1}^{n} X_j}{n} \tag{5.7}$$

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \left(\bar{X}\right)^2} \tag{5.8}$$

5.4.5. Generación de transacciones

A partir del conjunto de datos obtenido de la actividad anterior se procedió a la generación de datos transaccionales. Para esto, primeramente y como se especifica en la descripción del proceso de explotación de información de la sección 4.2.4, se calculó la distancia entre cada par de puntos mediante el uso de la herramienta para la generación de matrices de distancia a partir de datos vectoriales disponible en QGis (Qgis, 2020). Posteriormente, se filtraron aquellos pares de la relación de vecindad cuya separación fuese menor a 100 metros según la ortodrómica entre dos puntos calculada a partir de las coordenadas de los mismos. Cabe mencionar que debido a la escasa distancia entre los puntos, la curvatura de la geodésica resulta despreciable, por lo que la ortodrómica se aproxima a la distancia euclidiana.

Posteriormente, las relaciones de vecindad que cumplen con la condición estricta de vecindad fueron exportadas en un archivo CSV y usado como entrada de un *script* escrito en el lenguaje de programación Python. Este script hace uso de la aplicación *quick-cliques* (Strash, 2018) para la búsquedas de cliques máximos (mediante un enfoque híbrido en este caso) y traduce los cliques obtenidos según la descripción de los crímenes según el atributo *Descript*. De esta forma se obtuvieron 219 cliques de crímenes separados por menos de 100 metros de distancia, los cuales fueron plasmados en un archivo CSV donde cada fila del mismo posee una lista de las categorías de tipos de crímenes que se encuentran en cada uno de los vecindarios donde todos los crímenes son vecinos entre sí.

En la tabla 5.17 puede observarse la lista de todos los posibles valores que puede tomar el atributo en cuestión y la cantidad de instancias de cada uno de estos valores.

Tabla 5.17: Valores posibles del atributo “Descript” de la base de datos utilizada y la cantidad de instancias de cada clase dentro de las zonas calientes.

Valor	Cantidad
Attempted robbery chain store with bodily force	2
Attempted robbery on the street W/deadly weapon	2
Attempted robbery on the street with a gun	1
Attempted robbery on the street with bodily force	7
Attempted robbery with a gun	1
Attempted robbery with bodily force	9
Attempted stolen vehicle	1

Valor	Cantidad
Carjacking with a gun	2
Malicious Mischief, breaking windows	25
Malicious Mischief, breaking windows with BB gun	1
Malicious Mischief, graffiti	6
Malicious Mischief, street cars/buses	1
Malicious Mischief, vandalism	80
Malicious Mischief, vandalism of vehicles	81
Possession of Possession of base/rock cocaine	24
Possession of base/rock cocaine for sales	13
Possession of cocaine	3
Possession of cocaine for sales	2
Possession of controlled substance	13
Possession of controlled substance for sale	6
Possession of heroin	9
Possession of heroin for sales	3
Possession of marijuana	37
Possession of marijuana for sales	21
Possession of meth-amphetamine	32
Possession of meth-amphetamine for sale	6
Possession of methadone	2
Possession of narcotics paraphernalia	18
Possession of opiates	2
Possession of opiates for sales	1
Robbery of a chain store with a dangerous weapon	1
Robbery of a chain store with bodily force	6
Robbery of a commercial establishment, strongarm	7
Robbery of a residence with bodily force	1
Robbery on the street with a dangerous weapon	6
Robbery on the street with a gun	20
Robbery on the street with a knife	9
Robbery on the street, strongarm	114
Robbery, armed with a dangerous weapon	1
Robbery, armed with a gun	8

Valor	Cantidad
Robbery, armed with a knife	7
Robbery, bodily force	70
Sale of Base/Rock cocaine	11
Sale of marihuana	4
Sale of meth-amphetamine	2
Stolen and recovered vehicle	3
Stolen automobile	43
Stolen motorcycle	6
Stolen truck	10
Transportation of opiates	1

5.4.6. Generación de patrones

El archivo generado en la etapa anterior fue utilizado como entrada del algoritmo FP-Growth implementado en Rapidminer (RapidMiner, 2020) para la búsqueda de patrones de asociación debido a que no hay una clase de evento de referencia especificado para usar como atributo objetivo o *target*. En tal caso, la codificación debería realizarse mediantes atributos booleanos *doomies* y utilizar un algoritmo de clasificación para generación de reglas.

Se especificó como entrada del algoritmo un número mínimo de clases por patrón igual a 2, obteniendo dos patrones de tamaño 3 y ocho patrones de tamaño 2. Estos patrones así como el soporte de los mismos se exponen en la tabla 5.18. Cabe aclarar que bajo este enfoque cada clique puede estar conformado por múltiples instancias de una misma clase. En la conformación de los datos transaccionales se han omitido las clases duplicadas dentro de cada clique. La tabla anteriormente mencionada también incluye la cantidad de instancias únicas de cada clique dentro de la base de datos mediante la multiplicación de las instancias de cada clase en cada subgrafo y la posterior suma de estos cardinales para los 219 vecindarios considerados, por regla del producto y regla de la suma (Epp, 2011).

Tabla 5.18: Conjunto de crímenes que frecuentemente ocurren a menos de 100 metros de distancia uno de otro en las zonas calientes de la ciudad de San Francisco (USA). La columna de instancias especifica la cantidad de instancias únicas del patrón de co-localización.

Co-localización de crímenes	**Soporte**	**Instancias**
Robbery on the street, strongarm; Robbery on the street with a gun; Attemped robbery with bodily force	0.096	1863
Robbery on the street, strongarm; Possession of marijuana for sales; Robbery on the street with a knife	0.091	1560
Robbery, bodily force; Possession of meth-amphetamine	0.192	915
Robbery on the street, strongarm; Robbery on the street with a gun	0.151	735
Robbery, bodily force; Malicious mischief, Vandalism of vehicles	0.142	1420
Robbery on the street, strongarm; Possession of marijuana for sales	0.142	1522
Robbery on the street with a gun; Attemped robbery with bodily force	0.110	89
Robbery on the street, strongarm; Attemped robbery with bodily force	0.100	735
Robbery on the street, strongarm; Robbery on the street with a knife	0.091	525
Possession of marijuana for sales; Robbery on the street with a knife	0.091	163

5.4.7. Evaluación de los resultados

Primeramente, es necesario explicitar ciertas cuestiones relacionadas tanto con la base de datos como con la naturaleza de los resultados. En cuanto a la base de datos, existen múltiples clases de datos con una cantidad de instancias en la región explorada considerablemente reducida. Esto impacta en la posibilidad de encontrar un patrón frecuente que incluya a esta categoría de evento delictivo, lo cual es razonable. En cuanto a los resultados obtenidos, es necesario aclarar que la cantidad de instancias obtenidas por patrón no implica que haya una cantidad igual de hechos delictivos por clase, sino que existe esta cantidad de instancias del subgrafo considerado, calculado mediante propiedades de cálculo de cardinal de conjuntos. Estos cálculos pueden observarse en el archivo correspondiente del repositorio (Rottoli, 2020c). Asimismo, el soporte de la relación se calcula con base en la cantidad de vecindarios considerados, esto es, la cantidad de tuplas de la base de datos transaccional y por tanto la cantidad de cliques conformados.

En relación a los subgrafos obtenidos, el patrón detectado con mayor soporte que involucra tres clases de hechos delictivos distintas implica la ocurrencia de robos en la calle mediante intimidación y mediante el uso de armas de fuego e intento de robo mediante el uso de fuerza bruta. Esto ocurre con un soporte igual a 0.096, implicando 21 cliques que incluyen este patrón en la región explorada y 1863 instancias del mismo, lo cual significa un 11.6 % de la totalidad de posibles instancias de estas tres clases que ocurren en la zona considerada.

En segundo lugar, con un soporte igual a 0.091, se encuentra un subgrafo frecuente que involucra robos en la calle mediante intimidación, robos en la calle mediante el uso de cuchillos y casos de posesión de marihuana para ventas. Esto ocurre en 20 vecindarios o cliques, teniendo un total de 1560 instancias lo cual significa un 7.2 % de las posibles instancias del patrón con hechos delictivos de estas clases que ocurren en esta zona de la ciudad de San Francisco.

En cuanto a los patrones que involucran solo dos clases de hechos delictivos, el patrón con mayor soporte, igual a 0.192, determina la presencia de robos mediante el uso de fuerza bruta y la posesión de meta-anfetaminas a menos de 100 metros de distancia, lo cual se evidencia en cerca de 43 vecindarios. La cantidad de instancias del patrón presentes de forma vecina en el espacio, 915, significa un 40 % de las instancias posibles del mismo.

El segundo patrón implica la presencia de casos de robo mediante intimidación en la calle y casos de robo con armas de fuego con un soporte igual a 0.151 y una cantidad de instancias igual 735, lo cual se traduce como un 32,2 % del producto cartesiano entre las dos grupos de hechos delictivos en cuestión.

Además, con un soporte igual a 0.142, se evidencia la presencia a menos de 100 metros de separación de robos mediante fuerza bruta y vandalismo en vehículos, con un porcentaje de instancias de este patrón presente en zonas vecinas del sector considerado igual al 25,3 %. Además, con el mismo soporte se presenta un subgrafo frecuente que involucra robos mediante intimidación y la posesión de marihuana para ventas con un porcentaje de instancias de producto cartesiano presentes en las zonas calientes consideradas igual al 50.3 % con 1522 elementos en la relación.

Por último, los restantes tres patrones descubiertos son subgrafos de los dos patrones de tres clases mencionados al principio de este análisis. En el primer caso se ven involucrados robos en la calle mediante intimidación e intentos de robo mediante fuerza bruta con un soporte igual a 0.1 y un porcentaje de instancias relacionadas igual al 71.6 % de todas las posibles combinaciones de instancias de estas dos clases. La diferencia con el primer patrón de tres clases mencionado radica en la presencia de robos en la calle con armas de fuego, por lo que la cantidad de incidentes con esta modalidad parece ser menos probable.

Con un soporte igual a 0.091 se descubrieron los dos patrones restantes. Por un lado un patrón que involucra robos en la calle mediante intimidación y robos armados con cuchillos y por otro lado un patrón en el que se evidencia casos de posesión de marihuana para ventas y robos armados con cuchillos a menos de 100 metros de separación. La primera relación es un subconjunto con el 51.1 % de los posibles pares de eventos mientras que la segunda está compuesta por el 86.2 % de los pares de eventos posibles.

En resumen, los robos en la calle ya sean mediante intimidación, fuerza bruta o el uso de armas cortantes o de fuego se encuentran presente en todos patrones de colocalización. En todos los casos, el soporte de las relaciones no fue elevado, mas si lo fue el porcentaje de instancias que se vieron involucradas. Esto implica que estos patrones se encuentran mayormente localizados en pequeños espacios de 100 metros dentro de todo el área considerada, por lo que puede ser necesaria una segunda iteración mediante un agrupamiento más fino. El patrón con mayor cantidad de instancias involucradas fue el último. La posesión de marihuana para venta y el robo a mano

armada con cuchillos suelen suceder en ciertos radios pero se encuentran altamente relacionados. Un 86.2 % de las posibles vecindades entre las instancias de estas clases se encuentran modelados estos cliques descubiertos.

En cuanto a la posesión de meta-anfetaminas, mencionada en el tercer patrón, el mismo ocurre frecuentemente en escenarios donde es normal la presencia de robos mediante la fuerza. En este contexto, el soporte de la relación es relativamente elevado y también lo es la cantidad de pares ordenados involucrados en la relación. Por otro lado, otro hecho delictivo que consta en las reglas es el vandalismo sobre automóviles, el cual se detalla en el quinto renglón de la tabla 5.18. En este caso, si bien el soporte es alto y se encuentran en muchos cliques tanto este hecho como los robos mediante fuerza bruta, la cantidad de instancias involucradas no es demasiado elevada, por lo que es probable que muchos casos de ambas clases no se encuentren relacionados entre sí.

Finalmente, debido a que múltiples instancias de robos son discriminadas según el tipo de violencia en el que se incurre para su ejecución, puede ser necesaria una reclasificación de los datos para analizar otros posibles patrones según criterios más generales.

5.5. Opinión de expertos

En las secciones anteriores se han expuesto las demostraciones de funcionamiento de los artefactos propuestos: los procesos de explotación de información espacialmente referenciada. Estas demostraciones ilustran su implementación y su capacidad para la obtención de información a partir del conjunto de datos para la solución del dominio del problema.

Se ha ilustrado además el uso de relaciones espaciales difusas en los casos que se lo requiere, el uso de múltiples tipos de datos espaciales, el aseguramiento de la autocorrelación espacial mediante uso de algoritmos de agrupamiento, la consideración de la propiedad de heterogeneidad espacial considerando vecindarios útiles para el problema abordado y se ha mostrado de qué forma se pueden implementar los procesos haciendo uso de herramientas de sencillo uso y acceso, al ser todas de licencia libre.

En la presente sección se analizarán tres aspectos de los procesos que fueron tenidos en cuenta en el momento de su diseño: la utilidad de los mismos para la resolución de los problemas de negocio planteados, la facilidad para interpretar los resultados y

obtener conocimiento a partir de ellos, y la posibilidad de adaptarse a distintos escenarios.

Para esto, las siguientes secciones se organizan como sigue: La sección 5.5.1 presenta la metodología utilizada haciendo uso de conjuntos difusos sobre etiquetas lingüísticas para la agregación del juicio del grupo de expertos participantes; la sección 5.5.2 presenta las métricas consideradas y las preguntas que guían la opinión de los expertos para la obtención de los valores de las mismas; la sección 5.5.3 presenta el perfil de los expertos que participaron en la evaluación de los procesos; por último, la sección 5.5.4 presenta el modelo conformado y los resultados obtenidos.

5.5.1. Metodología

Para la evaluación de distintas dimensiones relacionados con los procesos de explotación de información espacialmente referenciada propuestos en esta tesis doctoral se utilizará el juicio de expertos, tal como se ha detallado con anterioridad al inicio de este capítulo. Para esto, a continuación se detallan los procedimientos destinados a la recolección de los datos en la sección 5.5.1.1 y a la agregación de los valores obtenidos mediante el uso de lógica difusa en la sección 5.5.1.2.

5.5.1.1. Recolección de datos

Cada uno de los expertos participantes fueron instruidos en el diseño de los artefactos y fueron ilustrados mediante los casos de demostración anteriormente listados. Posteriormente, se hizo uso de entrevistas con cada uno de ellos para recabar datos al respecto. Las entrevistas, según Johannesson and Perjons (2014), consisten en sesiones de comunicación entre el investigador y el experto respondiente, en el que el primero controla la interacción. En este contexto, si bien los cuestionarios resultan apropiados para recolectar información simple y rigurosa, las entrevistas resultan efectivas para recabar información sensible, elicitar emociones, actitudes, opiniones y experiencias. Las entrevistas permiten acceder a información más profunda sobre la experiencia del usuario en relación con determinado artefacto.

En este caso, se utilizó una estrategia no estructurada de conducción de entrevistas, utilizando una pregunta guía por cada dimensión evaluada y permitiendo que el entrevistado se explaye en sus apreciaciones al respecto con el propósito de no ser intrusivo

y sesgar estas opiniones.

Adicionalmente se solicitó al experto entrevistado que proporcione una etiqueta lingüística (entre el conjunto de etiquetas "Muy bajo", "Bajo", "Medio", "Alto" y "Muy Alto"), por cada dimensión evaluada y por cada proceso, que resuma su experiencia al respecto. Esta técnica adicional fue llevada acabo con el propósito de agregar las distintas opiniones que los expertos poseen sobre cada una de estas dimensiones haciendo uso de teoría de conjuntos difusos, particularmente usando operadores de agregación con ponderadores ordenados (OWA, *ordered weighted averaging operator*) (Yager, 1988) para conjugarlas en un único valor que exprese el nivel de satisfacción de los requerimientos de cada proceso. Este proceso se detalla en la sección siguiente.

5.5.1.2. Agregación de etiquetas difusas

En particular en esta tesis doctoral se ha optado por utilizar la propuesta de (Herrera-Viedma et al., 2006) por permitir agregar información difusa representada mediante etiquetas lingüísticas para múltiples dimensiones ponderadas según el nivel de importancia de cada una de ellas para el dominio evaluado (Herrera et al., 2009; Herrera and Martínez, 2000; Xu, 2008), y considerando la frecuencia con la que los valores son asignados por los expertos a cada dimensión. A continuación se resume y adapta dicha propuesta para el problema de evaluación de procesos de explotación de información. De aquí en más la descripción del método está tomada del artículo referido anteriormente.

Primeramente, se define una tupla de etiquetas o términos lingüísticos totalmente ordenados $S = (s_i), i \in \{0 \cdots \tau\}$ de forma tal que $s_i \geq s_j \iff i \geq j$, $|S|$ es impar y el término ubicado en el punto medio representa la expresión "aproximadamente 0.5". El resto de los términos son ubicados de forma simétrica alrededor de este valor y la semántica asociada con cada término está dada por el orden de la estructura. Por ejemplo, $S = (\text{Inexistente}, \text{Muy bajo}, \text{Bajo}, \text{Medio}, \text{Alto}, \text{Muy Alto}, \text{Total})$.

Se definen además de la siguiente forma un operador de negación $Neg(s_i)$, operadores de comparación $Max(s_i, s_j)$ y $Min(s_i, s_j)$, y dos operadores de agregación, en este caso un OWA inducido guiado por la mayoría (IOWA, *majority guided induced OWA*) que combina los valores numéricos de forma tal que se resuma en el resultado final el valor que represente la mayoría de valores similares a ser agregados.

$$Neg(s_i) = s_j | j = \tau - i \tag{5.9}$$

$$Max(s_i, s_j) = s_k | k = Max(i, j) \tag{5.10}$$

$$Min(s_i, s_j) = s_k | k = Min(i, j) \tag{5.11}$$

Definición 5.1 (Operador IOWA) *Se define el operador de agregación* $\Phi_W : (R \times R)^n \rightarrow R$ *con un vector de ponderación asociado* $W = (w_1, w_2, \cdots, w_n)$ *donde* $w_i \in [0, 1]$ *y* $\sum_i w_i = 1$*, de forma tal que dada una lista de pares de entrada* $\{(u_n, p_n)\}$ *esta es agregada según la siguiente expresión:*

$$\Phi_W((u_1, p_1), (u_2, p_2), \cdots, (u_n, p_n)) = \sum_{i=1}^{n} w_i * p_{\sigma(i)} \tag{5.12}$$

donde $\sigma : \{1, \cdots, \tau\} \rightarrow \{1, \cdots, \tau\}$ *es una permutación donde* $u_{\sigma(i)} \geq u_{\sigma(i+1)}, \forall i = 1, \cdots, \tau - 1$*. Esto es,* $(u_{\sigma(i)}, p_{\sigma(i)})$ *es el par donde* $u_{\sigma(i)}$ *es el i-ésimo mayor valor en el conjunto* $\{u_1, \cdots, u_n\}$*.*

En este caso, el reordenamiento del conjunto de valores p_i (variables argumentales) a ser agregados es inducido según un conjunto de valores u_i (variables de inducción de orden) asociados a cada uno de ellos.

Definición 5.2 (Operador MLIOWA) *Se define el operador IOWA guiado por la mayoría* $\Phi_Q : (R \times S)^n \rightarrow S$ *según la siguiente ecuación:*

$$\Phi_Q((u_1, p_1), (u_2, p_2), \cdots, (u_n, p_n)) = S_k \in S \tag{5.13}$$

donde

$$k = redondear(\sum_{i=1}^{n} w_i * ind(p_{\sigma(i)})) \tag{5.14}$$

y además cada uno de los componentes de esta definición se definen como sigue:

1. $\sigma : \{1, \cdots, n\} \rightarrow \{1, \cdots, n\}$ *es una permutación tal que* $u_{\sigma(i+1)} \geq u_{\sigma(i)}$*,* $\forall i = 1, \cdots, n - 1$*. Esto es,* $u_{\sigma(i)}$ *es el i-ésimo menor valor del conjunto* $\{u_i\}$*.*

2. $u_i = sup_i$, *esto es, el soporte del valor* p_i *obtenido como sigue, siendo* $\alpha \in \{1, \cdots, \tau\}$ *:*

$$sup_i = \sum_{j=1}^{n} sup_{ij} | sup_{ij} = \begin{cases} 1 & si \quad |ind(p_i) - ind(p_j)| \leq \alpha \\ 0 & \text{en caso contrario} \end{cases} \tag{5.15}$$

3. $ind : S \rightarrow \{1, \cdots, \tau\} | ind(s_i) = i$, *es el índice del término lingüístico ordenado.*

4. *Q es un cuantificador lingüístico difuso que representa el concepto de mayoría difusa en la agregación.*

5. $W = (w_1, \cdots, w_n), w_i \in [0, 1]$ *y* $\sum_i w_i = 1$ *es el vector de ponderadores o pesos, calculados a partir de los datos como sigue:*

$$w_i = Q\left(\frac{sup_{\sigma(i)}}{n}\right) \Big/ \sum_{j=1}^{n} Q\left(\frac{sup_{\sigma(j)}}{n}\right) \tag{5.16}$$

Debido a que se asume en el operador mencionado en la definición 5.2 que todos los valores lingüísticos o dimensiones son igualmente importantes, es necesario realizar una modificación al mismo para considerar estas diferencias en relevancia para el problema abordado.

Definición 5.3 (Operador MLIOWA ponderado) *Se define el operador MLIOWA Ponderado como una función* $\Phi_Q^I : (S \times S)^n \rightarrow S$ *definida según la siguiente ecuación:*

$$\Phi_Q^I((I_1, p_1), (I_2, p_2), \cdots, (I_n, p_n)) = \Phi_Q((u_1, p_1), (u_2, p_2), \cdots, (u_n, p_n)) \tag{5.17}$$

donde

$$u_i = \frac{sup_i + ind(I_i)}{2} \tag{5.18}$$

con

$$sup_i = \sum_{j=1}^{n} sup_{ij} | sup_{ij} = \begin{cases} 1 & si \quad |ind(I_i) - ind(I_j)| \leq \alpha \\ 0 & \text{en caso contrario} \end{cases} \tag{5.19}$$

donde I_i *es el grado lingüístico de importancia del valor* p_i *a ser agregado.*

Además,

$$w_i = Q\left(\frac{u_{\sigma(i)}}{n}\right) \Big/ \sum_{j=1}^{n} Q\left(\frac{u_{\sigma(j)}}{n}\right) \tag{5.20}$$

5.5.2. Dimensiones evaluadas

Tres dimensiones serán consideradas para dar respuesta a las preguntas de investigación planteadas en el capítulo 3 de la presente tesis doctoral: la utilidad, la interpretabilidad y la adaptabilidad de los procesos de explotación de información, cada una de las cuales serán detalladas en las secciones siguientes.

Además, como ya hemos mencionado, cada dimensión será medida haciendo uso de etiquetas lingüísticas que expresan el grado con el que el proceso de explotación de información cumple con la propiedad en cuestión. La selección de esta herramienta se encuentra motivada por la complejidad agregada de determinar valores exactos para propiedades subjetivas como lo son las mencionadas a partir de evaluaciones realizadas sobre la percepción de un determinado usuario basadas en la experiencia del mismo. El uso de una escala cualitativa y difusa permite a los expertos brindar su opinión de una forma subjetiva y basada en su experiencia. De utilizar una escala numérica continua, los expertos se verían obligados a decantarse por un valor que resultaría arbitrario mientras que el uso de una escala cualitativa acerca a los evaluadores a los conceptos que representan y permite capturar sus impresiones de una forma más confiable (Garg et al., 2013; Klir and Yuan, 1995).

Debido a que el método descripto en la sección anterior requiere el uso de un conjunto de etiquetas de cardinalidad impar, cinco etiquetas serán utilizadas: Muy bajo, Bajo, Medio, Alto, Muy Alto, ordenadas tal como fueron mencionadas.

5.5.2.1. Utilidad

La utilidad (Wieringa, 2014) se refiere a la percepción de los expertos para con la interacción entre el artefacto y determinados contextos (situables en la experiencia de los mismos) para resolver problemas de negocio determinados. Esta métrica hace referencia a la posibilidad que tiene cada proceso de explotación de información para brindar conocimiento que sea útil para la toma de decisiones y, por tanto, se relaciona directamente con la percepción de los usuarios.

La pregunta a la que debe responder el experto sobre cada proceso es: *¿Qué tan útil considera que es el proceso de explotación de información para resolver problemas de negocio que requiera este tipo de patrón de conocimiento?*

5.5.2.2. Interpretabilidad

La interpretabilidad de los resultados obtenidos por cada proceso tiene relación directa con el concepto de usabilidad. Se pretende determinar un valor para la inversa del esfuerzo necesario para derivar conclusiones a partir de la información recabada por las distintas actividades del proceso.

Para esto, el experto ha sido instruido en la semántica de cada salida de los procesos a través de los casos de demostración. El experto deberá determinar a partir de ejemplos prácticos si le es posible extraer información que pueda ser aplicable en un problema de negocio particular.

La pregunta que deberá responder el experto es: *¿En qué nivel las salidas del proceso facilitan la comprensión del conocimiento extraído a partir de los datos?*

5.5.2.3. Adaptabilidad

La adaptabilidad de los procesos de explotación de información será analizada teniendo en cuenta tres dimensiones: 1. los algoritmos y técnicas aplicables para cada actividad del proceso; 2. la cantidad de herramientas software que las implementan al día en el que esta tesis doctoral fue escrita, incluyendo *scripts* y lenguajes de programación; y 3. la posibilidad de integrar las distintas herramientas software utilizando las salidas de una como entrada de otra con el mínimo esfuerzo.

Para esto el experto deberá tener en cuenta un conjunto de posibles alternativas para cada actividad y las distintas posibilidades de implementación. En caso de no conocer ciertas herramientas para alguna actividad particular, el entrevistador le listará posibles alternativas.

Los expertos poseen experiencia en la implementación de procesos combinando distintas herramientas software por lo que determinarán el nivel de complejidad de la integración de las mismas, ya sea utilizando lenguajes de programación y librerías de minería de datos o las facilidades de importación y exportación de archivos que los entornos de descubrimiento de conocimiento brindan.

La pregunta general que deberá responder el experto es: *¿En qué nivel el proceso de explotación de información es fácilmente implementable utilizando herramientas tecnológicas actuales?*.

5.5.2.4. Opiniones

Cada uno de los expertos brinda además sus apreciaciones personales sobre los procesos de explotación de información en relación a las dimensiones evaluadas. Si bien estas consideraciones planteadas de forma textual motivan el juicio brindado mediante etiquetas, se analizan sus respuestas generales al resultar valiosas para la interpretación de los resultados de la agregación realizada y posibilitar identificar puntos de mejoras para futuras iteraciones del proceso de diseño.

5.5.3. Equipo de expertos

Los expertos consultados para la evaluación de las propuestas de esta tesis doctoral poseen un trasfondo variado relacionado con el uso de procesos de explotación de información, programación, sistemas de información geográficos, diseño de sistemas, entre otros.

A continuación se detalla un breve resumen de su experiencia profesional y académica que justifica su participación en este trabajo.

5.5.3.1. Experto 1

El primer experto participante en la evaluación de los procesos es Ingeniero en Sistemas de Información por la Universidad Tecnológica Nacional (Argentina). Su experiencia profesional lo sitúa en el desarrollo de sistemas de información geográficas (GIS) durante un período de cuatro años utilizando tecnologías ESRI - ArcGIS. El experto ha co-participado de la redacción de una guía de buenas prácticas para el desarrollo de software en JavaScript aplicado a GIS.

Su capacidad de manipulación de datos espacialmente referenciados y sus habilidades de programación hace que sea idóneo para evaluar aspectos de utilidad y adaptabilidad de los procesos aplicados. Adicionalmente, el profesional no posee demasiada experiencia relacionada a la minería de datos por lo que sus apreciaciones sobre la interpretabilidad de los procesos resulta valiosa.

5.5.3.2. Experto 2

La segunda experta posee un Doctorado en Ciencias Informáticas por la Universidad Nacional de La Plata (Argentina), y un Master en Ingeniería de Software, una Especialidad en Sistemas Expertos y un grado en Ingeniería en Sistemas de Información, todos por la Universidad Tecnológica Nacional (Argentina).

Sus estudios de posgrado se desarrollan en torno a los proyectos de explotación de información, la ingeniería de requisitos y la inteligencia de negocios. Por tales motivos, sus apreciaciones con respecto a la utilidad de los artefactos propuestos es de suma relevancia.

5.5.3.3. Experto 3

La tercer persona entrevistada acreditó estudios universitarios de posgrado que le otorgaron los grado de Doctorado en Ciencias informáticas en la Facultad de Informática de la Universidad Nacional de La Plata, Master en Ingeniería del Conocimiento por la Facultad de Informática de la Universidad Politécnica de Madrid, Licenciatura en Sistemas de Información y Analista de Sistemas por la Universidad Nacional de Luján.

La experta es investigadora en las áreas de Aprendizaje Automático, Sistemas Inteligentes, Explotación de Datos, Ingeniería del Conocimiento, entre otros. Además se ha desempeñado como docente en cursos y seminarios de grado y posgrado en distintas universidades nacionales y es co-autora de libros sobre Sistemas Expertos y Minería de Datos, así como directora de tesis de maestría y doctorado en estas áreas.

Este grado de experticia en el tema le confiere autoridad para dar su opinión en cuanto al diseño de los artefactos mencionados.

5.5.3.4. Experto 4

El cuarto experto consultado es Ingeniero en Sistemas de Información por la Universidad Tecnológica Nacional y se encuentra al momento de la realización de la evaluación finalizando sus estudios de maestría en Sistemas de Información con Orientación en Base de Datos por la misma institución.

El experto se ha desempeñado como Auxiliar de Primera Categoría y luego como profesor adjunto de la asignatura Tecnologías para la Explotación de Datos dictada

en quinto año de la carrera de Ingeniería en Sistemas de Información en la Universidad Tecnológica Nacional, por lo que posee conocimientos sobre minería de datos y sobre la implementación de procesos de explotación de información mediante el uso de herramientas varias. Adicionalmente se ha desempeñado profesionalmente como desarrollador de software y luego como líder de proyecto de desarrollo de software.

Este trasfondo hace que el profesional pueda determinar de forma apropiada las dimensiones de interpretabilidad y de adaptabilidad de los diseños propuestos.

5.5.3.5. Experto 5

El último experto entrevistado posee titulaciones de Doctorado en Ciencias Informáticas por la Universidad Nacional de La Plata (Argentina), Master en Ingeniería de Software por la Universidad Politécnica de Madrid (España), Magister en Ingeniería de Software y Especialista en Ingeniería de Sistemas Expertos, ambos por el Instituto Tecnológico de Buenos Aires (Argentina).

Sus áreas de investigación y docencia se centran en el uso de Tecnologías Inteligentes y Robótica Cognitiva. Además, es co-autor de artículos y libros centrados en la explotación de información y los sistemas expertos, así como también en el uso de matemáticas aplicadas.

Es por esto último que se aprecian sus sugerencias en cuanto al diseño de los artefactos, particularmente en relación al uso de estructuras de datos y algoritmos basados en grafos.

5.5.4. Modelo

Utilizando la misma nomenclatura especificada en la sección 5.5.1: "Metodología", en esta sección se procede a instanciar el modelo para la evaluación de los procesos de explotación de información propuestos en esta tesis doctoral considerando las opiniones de los expertos expuestas previamente.

5.5.4.1. Definiciones

Sea $S = (s_0 : \textit{Muy Bajo}, s_1 : \textit{Bajo}, s_2 : \textit{Medio}, s_3 : \textit{Alto}, s_4 : \textit{Muy Alto})$ la tupla de términos lingüísticos ordenados que se han usado para valorar el grado de satisfacción

Tabla 5.19: Grados de importancia de las dimensiones evaluadas de los procesos de explotación de información

d_i	d_1	d_2	d_3
$I(d_i)$	*Muy alto*	*Medio*	*Alto*

de las métricas definidas con anterioridad, el grado de importancia de las mismas y el grado de experticia de cada experto. Según la definición de S, el orden total implícito de los términos lingüísticos inducido por su orden es: $s_0 \prec s_1 \prec s_2 \prec s_3 \prec s_4$.

Sea $P = \{P_1, P_2, P_3, P_4\}$ el conjunto de procesos de explotación de información a ser evaluados, tal que P_1 : proceso de descubrimiento de grupos espaciales, P_2 : proceso de descubrimiento de anomalías espaciales, P_3 : proceso de descubrimiento de asociaciones espaciales y P_4 : proceso de descubrimiento de patrones de co-localización.

Sea $E = \{e_t^m\}$ el conjunto de evaluaciones de los expertos donde cada e_t^m es la evaluación realizada por el experto t del proceso p_m.

Sea $D = \{d_1, d_2, d_3\}$ el conjunto de las dimensiones a evaluar, tal que d_1 : Utilidad, d_2 : Interpretabilidad y d_3 : Adaptabilidad. Además, cada dimensión d_i está asociada con un grado de importancia $I(d_i) \in S$, tal como se especifica en la definición 5.3. Esta asociación se observa en la tabla 5.19, siendo la utilidad la dimensión que se ha considerado de mayor relevancia para el estudio, seguido de la adaptabilidad, debido a la naturaleza del dominio del problema, y por último la capacidad de interpretar los resultados.

Sea además el cuantificador lingüístico difuso Q=*"La mayoría"* definido de forma relativa con los parámetros (0.3, 0.7), y sea además $\alpha = 1$.

Sea entonces $\{d_i^{t,m}\}$ el conjunto de evaluaciones lingüísticas donde cada $d_i^{t,m} \in S$ es la evaluación de la dimensión i provista por el experto t en relación al proceso p_m, se busca generar un evaluación de calidad r^m de cada proceso m mediante la aplicación de los operadores MLIOWA. Esto se realiza en dos pasos según los autores originales del método. Estos dos pasos serán descriptos y desarrollados a continuación.

5.5.4.2. Agregación por dimensiones de calidad

En la tabla 5.20 se presenta el juicio de los expertos considerados para las tres dimensiones evaluadas según los expertos mencionados en la sección 5.5.3 mas no en

el mismo orden para anonimizar las respuestas.

En primer lugar se agregan las opiniones de los expertos agrupando sus juicios para cada dimensión de calidad considerada. Para esto, por cada experto e_t^m, su opinión individual r_t^m se calcula utilizando el operador ponderado MLIOWA Φ_Q^I como se especifica en la ecuación 5.17 para lo cual se deben obtener los valores u_i a partir de los grados de importancia anteriormente especificados según la ecuación 5.18. Estos valores se observan en la tabla 5.21 y especifican que el orden de las dimensiones consideradas es (d_2, d_1, d_3).

Luego, a partir de los valores inducidos obtenemos los ponderadores w_i utilizando la ecuación 5.20, siendo el vector resultante $W = (0.32, 0.34, 0.34)$. Los resultados parciales del cálculo se muestran en la tabla 5.22.

Este vector fue utilizado para ponderar las opiniones del juicio de experto (Tabla 5.23) para luego calcular el valor r_t^m para cada uno de los expertos considerados (Tabla 5.24).

5.5.4.3. Agregación por experto

Una vez agregadas las opiniones de los expertos para cada dimensión considerada en cada proceso se busca obtener un único valor r^m para cada proceso, agregando los valores obtenidos en el paso anterior, tal como se ha especificado anteriormente en la sección metodológica.

Para esto, se hace uso del operador de agregación MLIOWA como se ha definido en Def. 5.2, para lo cual es menester calcular en primera instancia el orden inducido de cada r_t^m haciendo uso de la ecuación 5.15 usando los datos de la tabla 5.24. Como resultado se obtiene que todos los expertos son ponderados de igual forma, debido a la cercanía de sus opiniones, tal como se evidencia en la tabla 5.25. Cabe aclarar que para la obtención de estos resultados se ha utilizado el parámetro $\alpha = 1$ y el mismo cuantificador difuso empleado en la sección previa.

En consecuencia, como se muestra en la tabla 5.26, se obtiene que todos los procesos de explotación de información evaluados cumplen con los requerimientos de utilidad, interpretabilidad y adaptabilidad en un nivel *Alto*, según los expertos participantes, validando de esta forma el diseño de los mismos.

Todos los cálculos auxiliares realizados para la obtención de los valores expuestos

Tabla 5.20: Juicio de experto para cada proceso de explotación de información con respecto a las dimensiones Utilidad, Interpretabilidad y Adaptabilidad utilizando etiquetas lingüísticas difusas.

			P_1		
	e_1	e_2	e_3	e_4	e_5
d_1	*Medio*	*Alto*	*Medio*	*Bajo*	*Alto*
d_2	*Muy Alto*	*Muy Alto*	*Alto*	*Muy Alto*	*Alto*
d_3	*Alto*	*Muy Alto*	*Alto*	*Alto*	*Alto*
			P_2		
	e_1	e_2	e_3	e_4	e_5
d_1	*Alto*	*Alto*	*Muy Alto*	*Muy Alto*	*Alto*
d_2	*Alto*	*Medio*	*Alto*	*Alto*	*Medio*
d_3	*Alto*	*Alto*	*Muy alto*	*Alto*	*Alto*
			P_3		
	e_1	e_2	e_3	e_4	e_5
d_1	*Muy Alto*	*Muy Alto*	*Alto*	*Alto*	*Muy Alto*
d_2	*Medio*	*Alto*	*Medio*	*Alto*	*Alto*
d_3	*Bajo*	*Medio*	*Alto*	*Medio*	*Alto*
			P_4		
	e_1	e_2	e_3	e_4	e_5
d_1	*Muy Alto*	*Alto*	*Alto*	*Alto*	*Muy Alto*
d_2	*Medio*	*Alto*	*Alto*	*Alto*	*Alto*
d_3	*Bajo*	*Medio*	*Medio*	*Medio*	*Alto*

Tabla 5.21: Valores de ordenamiento inducidos por el grado de importancia de las dimensiones evaluadas.

d_i	$I(d_i)$	Sup_i	$Ind(I(d_i))$	$u_i(I(d_i))$
d_1	*Muy Alto*	2	4	3
d_2	*Medio*	2	2	2
d_3	*Alto*	3	3	3

Tabla 5.22: Ponderadores obtenidos a partir del orden de las dimensiones evaluadas inducidos por su grado de importancia.

d_i	$\sigma(i)$	$u_{\sigma(i)}$	$Q(u_{\sigma(i)}/n)$	w_i
d_2	1	2	0.92	0.32
d_1	2	3	1	0.34
d_3	3	3	1	0.34

Tabla 5.23: (Izquierda) Índices de los términos lingüísticos obtenidos a partir del juicio de experto. (Derecha) Ponderación de los índices utilizando el vector W calculado con anterioridad.

	P_1				
	e_1	e_2	e_3	e_4	e_5
d_2	4	4	3	4	3
d_1	2	3	2	1	3
d_3	3	4	3	3	3

	P_2				
	e_1	e_2	e_3	e_4	e_5
d_2	3	2	3	3	2
d_1	3	3	4	4	3
d_3	3	3	4	3	3

	P_3				
	e_1	e_2	e_3	e_4	e_5
d_2	2	3	2	3	3
d_1	4	4	3	3	4
d_3	1	2	3	2	3

	P_4				
	e_1	e_2	e_3	e_4	e_5
d_2	2	3	3	3	3
d_1	4	3	3	3	4
d_3	1	2	2	2	3

	P_1				
	e_1	e_2	e_3	e_4	e_5
d_2	1.28	1.28	0.96	1.28	0.96
d_1	0.68	1.02	0.68	0.34	1.02
d_3	1.02	1.36	1.02	1.02	1.02

	P_2				
	e_1	e_2	e_3	e_4	e_5
d_2	0.96	0.64	0.96	0.96	0.64
d_1	1.02	1.02	1.36	1.36	1.02
d_3	1.02	1.02	1.36	1.02	1.02

	P_3				
	e_1	e_2	e_3	e_4	e_5
d_2	0.64	0.96	0.64	0.96	0.96
d_1	1.36	1.36	1.02	1.02	1.36
d_3	0.34	0.68	1.02	0.68	1.02

	P_4				
	e_1	e_2	e_3	e_4	e_5
d_2	0.64	0.96	0.96	0.96	0.96
d_1	1.36	1.02	1.02	1.02	1.36
d_3	0.34	0.68	0.68	0.68	1.02

Tabla 5.24: Valor del operador MLIOWA para cada proceso de explotación de información por cada experto consultado obtenidos al agregar las tres dimensiones evaluadas

	e_1	e_2	e_3	e_4	e_5
P_1	2.98 *Alto*	3.66 *Muy Alto*	2.66 *Alto*	2.64 *Alto*	3 *Alto*
P_2	3 *Alto*	2.68 *Alto*	3.68 *Muy Alto*	3.34 *Alto*	2.68 *Alto*
P_3	2.34 *Medio*	3 *Alto*	2.68 *Alto*	2.66 *Alto*	3.34 *Alto*
P_4	2.34 *Medio*	2.66 *Alto*	2.66 *Alto*	2.66 *Alto*	3.34 *Alto*

Tabla 5.25: Cálculo del orden y los valores de los ponderadores para los términos a ser agregados por cada proceso de explotación de información. Todos los procesos P_m poseen los mismos valores para cada una de los parámetros de I_Q

		e_1	e_2	e_3	e_4	e_5
P_m	$u_i = sup_i$	5	5	5	5	5
	$\sigma(i)$	1	2	3	4	5
	$Q(u_{\sigma(i)}/n)$	1	1	1	1	1
	w_i	0.20	0.20	0.20	0.20	0.20

Tabla 5.26: (Izquierda) Índices de los términos difusos resultantes de agregar las dimensiones evaluadas para cada experto y para cada uno de los procesos de explotación de información. (Derecha) Ponderación de los índices y aplicación del operador de agregación para cada uno de los procesos.

	e_1	e_2	e_3	e_4	e_5
P_1	3	4	3	3	3
P_2	3	3	4	3	3
P_3	2	3	3	3	3
P_4	2	3	3	3	3

	e_1	e_2	e_3	e_4	e_5	$\sum$	Etiqueta
P_1	0.6	0.8	0.6	0.6	0.6	3.2	*Alto*
P_2	0.6	0.6	0.8	0.6	0.6	3.2	*Alto*
P_3	0.4	0.6	0.6	0.6	0.6	2.8	*Alto*
P_4	0.4	0.6	0.6	0.6	0.6	2.8	*Alto*

en esta sección se encuentran detallados en el anexo B de este documento.

5.6. Discusión

A partir de los comentarios realizados por los expertos durante la validación de los procesos de explotación de información (Rottoli, 2020e), se extraen las opiniones que se expresan a continuación.

Los procesos de explotación de información anteriormente evaluados poseen un nivel agregado *Alto* de cumplimiento de las dimensiones de utilidad (que considera su aplicación para la resolución de problemas en dominios particulares), la adaptabilidad (referida a la capacidad de adaptarlos haciendo uso de tecnologías específicas) y a la interpretabilidad de los resultados (relacionados con la forma en la que son extraídas las relaciones implícitas a partir de los datos). No obstante se deben hacer ciertas apreciaciones en cuanto a los comentarios particulares de los expertos para con los mismos.

En primer lugar, el proceso de explotación de información para el descubrimiento de grupos espaciales ha sido apreciado con niveles medio y bajo en cuanto a la utilidad

se refiere por tres de los cinco profesionales. Según los expertos, el proceso de explotación de información no resulta demasiado complejo, es intuitivo y no aporta demasiado valor agregado en relación a lo que se realiza actualmente en la práctica. Posee demasiada similitud con el proceso de explotación de información para el descubrimiento de reglas de pertenencia a grupos propuesto por (Britos, 2008). Sin embargo, resulta interesante la posibilidad de distinguir entre las distintas problemáticas (determinación de grupos, regiones, y zonas calientes) y poder aplicar la misma estrategia para las tres secciones, principalmente en cuanto a regionalización se refiere ya que no es una categoría demasiado usada y que posee una alta posible transferencia a la industria.

El proceso de explotación de información para el descubrimiento de anomalías espaciales, por el contrario, fue apreciado positivamente por la gran mayoría de los expertos. La dimensión más pobremente evaluada fue la correspondiente a la interpretabilidad de los resultados. Según los expertos, múltiples salidas pueden ser difíciles de integrar, sobre todo si el nivel de abstracción es tan alto que resulta difícil extrapolarlo a un caso concreto. Ante esto, el caso de demostración presentado resulta de ayuda considerable.

Luego, el proceso de explotación de información para el descubrimiento de asociaciones espaciales obtuvo una mayor disparidad en las opiniones de los expertos. Resulta interesante la apreciación común sobre la complejidad agregada de la lógica difusa, cuestión que pudo haber tenido mayor protagonismo por el caso de demostración presentado. Concretamente, la evaluación de los resultados puede resultar una actividad nada sencilla de realizar. En relación a la utilidad del proceso, los expertos opinan que la misma es elevada, siendo altamente aplicable. La adaptabilidad del método tiende a ser vista de forma distinta por cada uno de los consultados. Dos expertos opinan que se encuentra en un nivel medio, dos en nivel alto, y uno en nivel bajo. Esto se debe a la poca presencia de algoritmos para el minado de subgrafos frecuentes en las *suits* de software para la explotación de información y a la necesidad de recurrir generalmente a métodos programáticos para la extracción de las relaciones espaciales. Sin embargo, distintas herramientas y algoritmos de terceras partes pueden ser aprovechados para estas tareas.

En relación con la interpretabilidad del proceso de descubrimiento de asociaciones espaciales, los expertos opinan que las estructuras gráficas son comprensibles y resulta intuitiva la traducción entre grafos y reglas. Sin embargo, el inconveniente más frecuente es que no se encuentran habituados a interpretar las métricas de los algoritmos

de descubrimiento de subgrafos frecuentes, como la tasa de comprensión, por lo que requieren calcular niveles de soporte y confianza como sucede con las tradicionales reglas de asociación de datos relacionales.

Por último, el proceso de explotación de información para el descubrimiento de patrones de co-localización ha sido visto de forma similar al anterior por los evaluadores. Se percibe una alta utilidad por parte de todos los profesionales al ser un patrón fácilmente comprensible y trasladable a situaciones concretas. La interpretabilidad de los resultados resulta alta también, principalmente por estar familiarizados con las asociaciones de datos sobre bases de datos transaccionales tradicionales. Por último, la adaptabilidad del proceso es percibida en un nivel medio por tres expertos, mientras que los dos restantes lo perciben baja y alta respectivamente. Esta apreciación se refiere a que, por lo que respecta al conocimiento de los expertos, no existe demasiada variedad de artefactos software para la búsqueda de cliques máximos.

En resumen, los expertos han establecido los niveles de utilidad, interpretabilidad y adaptabilidad de los procesos de explotación de información propuestos en esta tesis doctoral, determinando que los mismos cumplen con los requerimientos especificados en un nivel suficientemente aceptable. Valiosos comentarios con respecto a cada uno de ellos han sido rescatados de su participación en esta evaluación, los cuales han de ser tenidos en consideración en las iteraciones de mejora subsiguientes.

CAPÍTULO 6

CONCLUSIÓN Y TRABAJOS FUTUROS

"Things have changed, and will change more (...)"

Neil Gaiman, A Midsummer Night's Dream

En este capítulo se concluye el desarrollo de esta tesis doctoral mediante los aportes significativos de la misma al estado del arte, descriptos en la sección 6.1, y las líneas de investigación pendientes a ser exploradas en futuros trabajos científicos en la sección 6.2.

6.1. Aportes de la tesis

En la presente tesis doctoral se han presentado cuatro procesos de explotación de información espacialmente referenciada que sistematiza la obtención de conocimiento a partir de este tipo de datos de entrada considerando las particularidades asociadas a la contextualización de los mismos.

En relación al primer y tercer objetivo específico planteado se han diseñado los siguientes procesos de explotación de información, dando respuesta a las preguntas de investigación P1, P2:

- Un proceso para el descubrimiento de grupos espaciales que permite distinguir entre grupos, regiones y zonas calientes utilizando para cada caso algoritmos

propuestos de distinta naturaleza. Posteriormente los grupos son descriptos a partir de los atributos no espaciales de los datos.

- Un proceso para el descubrimiento de anomalías espaciales que posibilita la búsqueda de datos espaciales desviados en relación a sus contextos locales, la descripción del comportamiento de los datos normales según los atributos no espaciales de los mismos y la búsqueda de grupos de datos espaciales con comportamiento anómalo.

- Un proceso para el descubrimiento de asociaciones espaciales que permite la búsqueda de estas regularidades en contextos locales, la consideración de múltiples tipos de relaciones espaciales simultáneas y la posibilidad de utilizar descripciones difusas de las mismas. Se utiliza para esto teoría de grafos y algoritmos de minería de grafos bien establecidos en el estado del arte.

- Un proceso para el descubrimiento de patrones de co-localización, que también permite la búsqueda de estas regularidades en contextos locales y su extensión a múltiples tipos de relaciones. Se utilizan, como en el caso anterior, minería de grafos para la obtención de transacciones que son minadas con algoritmos de minería de asociaciones tradicionales.

.

En cuanto al segundo objetivo específico planteado, se destaca que cada uno de los procesos anteriormente mencionados cumplen con las siguientes características generales, respondiendo en consecuencia la pregunta de investigación P3:

- Utilizan datos vectoriales como entrada, pudiendo ser estos datos puntuales, líneas o polígonos. Por este motivo resultan adaptables a las distintas necesidades de modelado del dominio del problema.

- Consideran, cuando es necesario, la posibilidad de representar múltiples relaciones espaciales, generalmente haciendo uso de teoría de grafos. Estas relaciones a su vez pueden tener naturaleza rígida o difusa, posibilitando el análisis de los grados de pertenencia de los elementos de esta relación en las etapas de evaluación de resultados.

- Consideran el fenómeno de la autocorrelación espacial si es requerido, mediante la utilización de algoritmos de agrupamiento basados en estadísticas, densidades, entre otros.

- Consideran el fenómeno de la heterogeneidad espacial al permitir dividir el espacio en sectores de interés para la toma de decisiones estratégicas. Esta división puede realizarse según criterios del experto, según conocimiento del dominio o mediante la utilización de algoritmos de agrupamiento de datos espaciales.

Por último, en cuanto a la implementación de los procesos establecida en el tercer objetivo específico y relacionado con la pregunta de investigación P4, se destaca que:

- Los procesos de explotación de información diseñados sugieren la aplicación de tipos de algoritmos algoritmos que han sido probados en la literatura, por lo que los resultados son fiables. Estos procesos sistematizan el uso de los mismos para la consideración de las características asociadas a la naturaleza espacial de los datos, mencionadas con anterioridad.

- En todos los casos se utilizan algoritmos bien establecidos y que resultan posibles de ser encontrados ya sea en las herramientas software disponibles en el mercado, herramientas de código abierto o librerías de programación. Esto ocurre tanto en el caso de los algoritmos de minería de datos relacionales, minería de datos espaciales, e incluso minería de grafos.

- Los procesos son independientes de los algoritmos a ser aplicados y de las tecnologías informáticas actuales, por lo que pueden ser implementados en múltiples plataformas haciendo uso de variedad de herramientas disponibles para uso comercial o incluso mediante librerías de programación.

- Son factibles de ser automatizados si se lo requiere.

Tal como se propone en el *framework* de ciencia de diseño (Wieringa, 2014), los artefactos propuestos en esta tesis doctoral fueron demostrados haciendo uso de casos de ejemplo y posteriormente evaluados mediante técnicas descriptivas utilizando el juicio de expertos. Se ha establecido que los mismos cumplen satisfactoriamente con los requerimientos especificados de utilidad, interpretabilidad de los resultados y de adaptabilidad a diferentes contextos y herramientas.

6.2. Futuras líneas de investigación

Los procesos de explotación de información presentados en esta tesis doctoral constituyen un primer conjunto de procedimientos para la extracción de conocimiento de forma adaptable y haciendo uso de técnicas accesibles, sin perder aspectos relevantes relacionados con la puesta en contexto de los datos en un espacio determinado.

En el ciclo de diseño de estos procesos han surgido inquietudes que no han sido consideradas en el alcance de este trabajo, por lo que restan ser encaradas en futuros trabajos. Estas cuestiones se enumeran a continuación:

- Los procesos de explotación de información diseñados permiten extraer cuatro tipos de regularidades a partir de los datos espaciales, mas existe una amplia variedad de patrones que han sido relegados para futuras iteraciones del proceso de diseño. En particular, resulta necesario abordar la clasificación de datos espacialmente referenciados, actividad de sumo interés para la inteligencia de negocio enfocada principalmente en los datos de naturaleza raster o basada en grillas, como lo son las imágenes censadas remotamente.

- En relación al punto anterior, si bien existe una forma de pasaje entre un tipo de representación raster de datos espaciales a un tipo de representación vectorial, se ha mencionado que tal conversión conlleva generalmente pérdida de información, por lo que se valora la existencia de procesos de explotación de información que operen con esta clase de datos de entrada.

- En otro ángulo, los artefactos diseñados pecan de pérdida de especificidad al haberse diseñado concienzudamente para poseer un grado de generalidad elevado. Se busca abstraer los problemas de dominio por lo cual ante escenarios con particularidades demasiado específicas resulta necesario el uso de los algoritmos del estado del arte especialmente diseñado para abordarlas, cuestión que ha sido identificada como parte de las problemáticas que dieron origen a este trabajo pero que no dejan de ser necesarias en otros contextos de aplicación. Se requiere un estudio sistemático de la bibliografía que mapee estos problemas particulares con los algoritmos que han sido propuestos.

- Asimismo, los procesos presentados en este trabajo consideran cuestiones como la heterogeneidad espacial, al especificar localidades de interés, y la posibilidad de considerar la autocorrelación espacial, ya sea con el uso de algoritmos de

clustering o con algoritmos basados en estadísticas. Sin embargo, existen otros fenómenos o características que no han sido exploradas y que ante ciertos escenarios pueden resultar de interés. Tales son los caso de la *Spatial Anisotropy* y de las múltiples escalas y resoluciones de datos (Jiang and Shekhar, 2017), si bien este último factor puede ser considerado al utilizar relaciones topológicas de inclusión, recalculando por cada nivel las relaciones de interés. Sin embargo, actualmente realizar esto puede incurrir en costos extras tanto de extracción de relaciones, modelado y de búsqueda de regularidades.

- Luego, queda pendiente una nueva iteración sobre los procesos presentados para considerar las opiniones de los expertos. Entre estas opiniones se destaca la falta de métricas comúnmente utilizadas, sobre todo en relación a los métodos para la obtención de asociaciones y de patrones de co-localización, al no utilizarse algoritmos que hayan sido diseñados para esto. Si bien en algunos casos resulta factible el cálculo de estas medidas, en otros no resulta posible, siendo un ejemplo particular aquellos en los que se usan métodos no basados en enfoques transaccionales, por lo que métricas como el soporte y la confianza de reglas deberían ser reformulados.

- Por último, estos procesos eventualmente pueden quedar obsoletos al hacerse más común la presencia de algoritmos especializados para la obtención de conocimiento en datos espacialmente referenciados en herramientas software, por lo que se requerirá oportunamente una actualización de los mismos para considerar estas alternativas dentro de las actividades que los conforman.

REFERENCIAS

Abdelhamid, E., Canim, M., Sadoghi, M., Bhattacharjee, B., Chang, Y.-C., and Kalnis, P. (2017). Incremental frequent subgraph mining on large evolving graphs. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2710–2723. https://doi.org/10.1109/TKDE.2017.2743075.

Acatitla Romero, E. and Urbina Alonso, J. (2017). El uso de redes complejas en economía: alcances y perspectivas. *INTERdisciplina*, 5(12):9–22.

Adams, J. and Strother-Adams, P. (2001). *Dealing with diversity: The anthology*. Kendall/Hunt Pub.

Aggarwal, C. C. (2015). *Data mining: the textbook*. Springer.

Aggarwal, C. C., Wang, H., et al. (2010). *Managing and mining graph data*, volume 40. Springer. https://doi.org/10.1007/978-1-4419-6045-0.

Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM. https://doi.org/10.1145/170035.170072.

Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499.

An-yang, W. and Wei-dong, Z. (2005). Discovery of spatial association rules based on multiple minimum supports. *Computer Applications*, 25(9):2171–2174.

Anselin et al. (2020). Geoda. https://geodacenter.github.io.

Appice, A., Ceci, M., Lanza, A., Lisi, F. A., and Malerba, D. (2003). Discovery of spatial association rules in geo-referenced census data: A relational mining approach. *Intelligent Data Analysis*, 7(6):541–566. https://doi.org/10.3233/IDA-2003-7604.

Arribas-Bel, D. (2020). Airbnb data at middle layer super output area (msoa) level. CC0 1.0 Universal License. http://darribas.org/gds15/content/labs/data/ilm_abb.geojson (25 de junio de 2020).

Assunção, R. M., Neves, M. C., Câmara, G., and da Costa Freitas, C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20(7):797–811. https://doi.org/10.1080/13658810600665111.

Atluri, G., Karpatne, A., and Kumar, V. (2018). Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys (CSUR)*, 51(4):83. https://doi.org/10.1145/3161602.

Azevedo, A. I. R. L. and Santos, M. F. (2008). Kdd, semma and crisp-dm: a parallel overview. *IADS-DM*. http://hdl.handle.net/10400.22/136.

Babu, H. P., Selvaraj, J., Ramachandran, S., Marimuthu, P. D., and Somanathan, B. (2015). Spatial data mining using association rules and fuzzy logic for autonomous exploration of geo-referenced cancer data in western tamilnadu, india. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 4(1):23. https://doi.org/10.1007/s13721-015-0094-1.

Bai, H., Ge, Y., Wang, J., Li, D., Liao, Y., and Zheng, X. (2014). A method for extracting rules from spatial data based on rough fuzzy sets. *Knowledge-Based Systems*, 57:28–40. https://doi.org/10.1016/j.knosys.2013.12.008.

Bansal, R., Gaur, N., and Singh, S. N. (2016). Outlier detection: applications and techniques in data mining. In *2016 6th International Conference-Cloud System and Big Data Engineering (Confluence)*, pages 373–377. IEEE. https://doi.org/10.1109/CONFLUENCE.2016.7508146.

Barabási, A.-L. (2003). Linked: The new science of networks. https://doi.org/10.1353/pbm.2004.0030.

Barabási, A.-L. et al. (2016). *Network science*. Cambridge university press.

Bianco, S. (2016). Análisis comparativo de algoritmos de minería de subgrafos frecuentes. *Revista Latinoamericana de Ingenieria de Software*, 4(2):111–142.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104. https://doi.org/10.1145/342009.335388.

Brimicombe, A. J. (2007). A dual approach to cluster discovery in point event data sets. *Computers, environment and urban systems*, 31(1):4–18. https://doi.org/10.1016/j.compenvurbsys.2005.07.004.

Britos, P. V. (2008). *Procesos de explotación de información basados en sistemas inteligentes*. PhD thesis, Facultad de Informática, Universidad Nacional de La Plata.

Brittingham, A. and de la Cruz, G. P. (2004). Ancestry 2000. oficina nacional del censo. En inglés. https://www.census.gov/prod/2004pubs/c2kbr-35.pdf (14 de agosto de 2020).

Cai, Q., He, H., and Man, H. (2013). Spatial outlier detection based on iterative self-organizing learning model. *Neurocomputing*, 117:161–172.

Celik, M., Kang, J. M., and Shekhar, S. (2007). Zonal co-location pattern discovery with dynamic parameters. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 433–438. IEEE.

Census.gov (2013). States, counties, and statistically equivalent entities (chapter 4). https://www2.census.gov/geo/pdfs/reference/GARM/Ch4GARM.pdf, (14 de agosto de 2020).

Center for Spatial Data Science (2017). San francisco crime incidents (2012). https://geodacenter.github.io/data-and-lab/SFcrimes_vars/.

Chang, K.-T. (2006). *Introduction to geographic information systems*. McGraw-Hill Higher Education Boston.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (1999). The crisp-dm user guide. In *4th CRISP-DM SIG Workshop in Brussels in March*, volume 1999.

Chawla, S. and Sun, P. (2006). Slom: a new measure for local spatial outliers. *Knowledge and Information Systems*, 9(4):412–429. https://doi.org/10.1007/s10115-005-0200-2.

Chen, H., Schroeder, J., Hauck, R. V., Ridgeway, L., Atabakhsh, H., Gupta, H., Boarman, C., Rasmussen, K., and Clements, A. W. (2003). Coplink connect: information and knowledge management for law enforcement. *Decision support systems*, 34(3):271–285. https://doi.org/10.1016/S0167-9236(02)00121-5.

Chen, J. et al. (2008). An algorithm about association rule mining based on spatial autocorrelation. *Intern. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 37(PART B6B):99–106.

Clementini, E., Di Felice, P., and van Oosterom, P. (1993). A small set of formal topological relationships suitable for end-user interaction. In Abel, D. and Chin Ooi, B., editors, *Advances in Spatial Databases*, pages 277–295. Springer Berlin Heidelberg. http://doi.org/10.1007/3-540-56869-7_16.

Clementini, E., Sharma, J., and Egenhofer, M. J. (1994). Modelling topological spatial relations: Strategies for query processing. *Computers & Graphics*, 18(6):815 – 822. https://doi.org/10.1016/0097-8493(94)90007-8.

Cressie, N. (1992). Statistics for spatial data. *Terra Nova*, 4(5):613–617. https://doi.org/10.1111/j.1365-3121.1992.tb00605.x.

Cunjin, X. and Xiaohan, L. (2017). Novel algorithm for mining enso-oriented marine spatial association patterns from raster-formatted datasets. *ISPRS International Journal of Geo-Information*, 6(5):139. https://doi.org/10.3390/ijgi6050139.

De Maesschalck, R., Jouan-Rimbaud, D., and Massart, D. L. (2000). The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18. https://doi.org/10.1016/S0169-7439(99)00047-7.

Deepak, P. (2016). Anomaly detection for data with spatial attributes. In *Unsupervised Learning Algorithms*, pages 1–32. Springer. https://doi.org/10.1007/978-3-319-24211-8_1.

Dharni, C. and Bnasal, M. (2013). An improvement of dbscan algorithm to analyze cluster for large datasets. In *2013 IEEE International Conference in MOOC, Innovation and Technology in Education (MITE)*, pages 42–46. IEEE. https://10.1109/MITE.2013.6756302.

Du, S., Wang, Q., and Guo, L. (2010). Modeling the scale dependences of topological relations between lines and regions induced by reduction of attributes. *International Journal of Geographical Information Science*, 24(11):1649–1686. https://doi.org/10.1080/13658811003591672.

Epp, S. S. (2011). *Discrete mathematics with applications*. Cengage learning.

Ernst, M. and Haesbroeck, G. (2017). Comparison of local outlier detection techniques in spatial multivariate data. *Data mining and knowledge discovery*, 31(2):371–399. https://doi.org/10.1007/s10618-016-0471-0.

Esfandani, G. and Abolhassani, H. (2010). Msdbscan: multi-density scale-independent clustering algorithm based on dbscan. In *International Conference on Advanced Data Mining and Applications*, pages 202–213. Springer. https://doi.org/10.1007/978-3-642-17316-5_19.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Estivill-Castro, V. and Houle, M. E. (2001). Robust distance-based clustering with applications to spatial data mining. *Algorithmica*, 30(2):216–242.

Euler, L. (1741). Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, pages 128–140.

Faridi, M., Verma, S., and Mukherjee, S. (2015). Association rule mining for ground water and wastelands using apriori algorithm: case study of jodhpur district. *International Journal of Advanced Research in Computer Science and Software Engineering*, 5(6):751–758.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37. https://doi.org/10.1609/aimag.v17i3.1230.

Ferreira, J. E., Takai, O. K., and Pu, C. (2005). Integration of business processes with autonomous information systems: a case study in government services. In *Seventh IEEE International Conference on E-Commerce Technology (CEC'05)*, pages 471–474. IEEE. https://doi.org/10.1109/ICECT.2005.56.

Gandhi, N. and Armstrong, L. J. (2016). A review of the application of data mining techniques for decision making in agriculture. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, pages 1–6. IEEE.

García-Martínez, R., Britos, P., Pesado, P., Bertone, R., Pollo-Cattaneo, M. F., Rodríguez, D., Pytel, P., and Vanrell, J. (2011). Towards an information mining engineering. *En Software Engineering, Methods, Modeling and Teaching. Sello Editorial Universidad de Medellín*, pages 83–99.

García-Martínez, R., Britos, P., and Rodríguez, D. (2013). Information mining processes based on intelligent systems. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 402–410. Springer. https://doi.org/10.1007/978-3-642-38577-3_41.

Garg, R., Sharma, K., Nagpal, C., Garg, R., Garg, R., and Kumar, R. (2013). Ranking of software engineering metrics by fuzzy-based matrix methodology. *Software Testing, Verification and Reliability*, 23(2):149–168.

Getis, A. and Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. In *Perspectives on spatial data analysis*, pages 127–145. Springer.

Getis, A. and Ord, J. K. (2010). The analysis of spatial association by use of distance statistics. In *Perspectives on spatial data analysis*, pages 127–145. Springer.

Gilkey, J. (2019). Graph theory and data science a topic intro with the bridges of königsberg. https://towardsdatascience.com/graph-theory-and-data-science-ec95fe2f31d8.

Golmohammadi, J., Xie, Y., Gupta, J., Li, Y., Cai, J., Detor, S., and Shekhar, S. (2018). An introduction to spatial data mining. Technical report, Technical report TR 18-013). Retrieved from https://www. cs. umn. edu/sites ....

Grinstein, G. G. and Ward, M. O. (2002). Introduction to data visualization. *Information visualization in data mining and knowledge discovery*, 1:21–45.

Guo, D. (2008). Regionalization with dynamically constrained agglomerative clustering and partitioning (redcap). *International Journal of Geographical Information Science*, 22(7):801–823. https://doi.org/10.1080/13658810701674970.

Han, J., Kamber, M., and Tung, A. K. (2001). Spatial clustering methods in data mining. *Geographic data mining and knowledge discovery*, pages 188–217.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without candidate generation. *ACM sigmod record*, 29(2):1–12.

He, Y., Tan, H., Luo, W., Feng, S., and Fan, J. (2014). Mr-dbscan: a scalable mapreduce-based dbscan algorithm for heavily skewed data. *Frontiers of Computer Science*, 8(1):83–99. https://doi.org/10.1007/s11704-013-3158-3.

Herrera, F., Alonso, S., Chiclana, F., and Herrera-Viedma, E. (2009). Computing with words in decision making: foundations, trends and prospects. *Fuzzy Optimization and Decision Making*, 8(4):337–364. https://doi.org/10.1007/s10700-009-9065-2.

Herrera, F. and Martínez, L. (2000). A 2-tuple fuzzy linguistic representation model for computing with words. *IEEE Transactions on fuzzy systems*, 8(6):746–752.

Herrera-Viedma, E., Pasi, G., Lopez-Herrera, A. G., and Porcel, C. (2006). Evaluating the information quality of web sites: A methodology based on fuzzy computing with words. *Journal of the American Society for Information Science and Technology*, 57(4):538–549.

Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). Design science in information systems research. *MIS quarterly*, pages 75–105.

Hodge, V. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9.

Holder, L. (2020). Subdue graph miner. https://github.com/holderlb/Subdue.

Holder, L. B., Cook, D. J., and Djoko, S. (1994). Substructure discovery in the subdue system. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, AAAIWS'94, page 169–180. AAAI Press. http://doi.org/10.5555/3000850.3000868.

IDESA (2017). Precipitación media mensual - noviembre. http://geoportal.idesa.gob.ar/layers/geonode:_11_pp.

Jiang, Z. and Shekhar, S. (2017). Spatial big data science. *Schweiz: Springer International Publishing AG*. https://doi.org/10.1007/978-3-319-60195-3.

Johannesson, P. and Perjons, E. (2014). *An introduction to design science*. Springer.

Kavitha, D., Prasad, V. K., and Murthy, J. (2016). Finding frequent subgraphs in a single graph based on symmetry. *International Journal of Computer Applications*, 146(11):5–8.

Kim, J. and Cho, J. (2019). Delaunay triangulation-based spatial clustering technique for enhanced adjacent boundary detection and segmentation of lidar 3d point clouds. *Sensors*, 19(18):3926.

Kim, S. K., Kim, Y., and Kim, U. (2011). Maximal cliques generating algorithm for spatial co-location pattern mining. In *FTRA International Conference on Secure and Trust Computing, Data Management, and Application*, pages 241–250. Springer. https://doi.org/10.1007/978-3-642-22339-6_29.

Kim, S. K., Lee, J. H., Ryu, K. H., and Kim, U. (2014). A framework of spatial co-location pattern mining for ubiquitous gis. *Multimedia tools and applications*, 71(1):199–218. https://doi.org/10.1007/s11042-012-1007-2.

Klir, G. and Yuan, B. (1995). *Fuzzy sets and fuzzy logic*, volume 4. Prentice hall New Jersey.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69. https://doi.org/10.1007/BF00337288.

Kuna, H. D., García Martínez, R., and Villatoro, F. (2009). Procedimientos de la explotación de información para la identificación de datos faltantes, con ruido e inconsistentes. In *XI Workshop de Investigadores en Ciencias de la Computación*.

Ladner, R., Petry, F. E., and Cobb, M. A. (2003). Fuzzy set approaches to spatial data mining of association rules. *Transactions in GIS*, 7(1):123–138. https://doi.org/10.1111/1467-9671.00133.

Lazzari, L., Casparri, M., Loisso, G., and Mouliá, P. (2001). Los conjuntos borrosos y su aplicación a la programación lineal. *Facultad de Ciencias Económicas, Universidad de Buenos Aires, Buenos Aires*.

Lee, I. (2004). Mining multivariate associations within gis environments. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 1062–1071. Springer. https://doi.org/10.1007/978-3-540-24677-0_109.

Legendre, P. (1993). Spatial autocorrelation: trouble or new paradigm? *Ecology*, 74(6):1659–1673. https://doi.org/10.2307/1939924.

Li, D., Wang, S., and Li, D. (2015). *Spatial data mining*. Springer.

Li, D., Wang, S., Yuan, H., and Li, D. (2016). Software and applications of spatial data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 6(3):84–114. https://doi.org/10.1002/widm.1180.

Li, H., Calder, C. A., and Cressie, N. (2007). Beyond moran's i: testing for spatial dependence based on the spatial autoregressive model. *Geographical Analysis*, 39(4):357–375.

Lisi, F. and Malerba, D. (2002). Spada: A spatial association discovery system. *WIT Transactions on Information and Communication Technologies*, 28.

Lu, C.-T., Chen, D., and Kou, Y. (2004). Multivariate spatial outlier detection. *International Journal on Artificial Intelligence Tools*, 13(04):801–811.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Maiti, S. and Subramanyam, R. (2018). Mining co-location patterns from distributed spatial data. *Journal of King Saud University-Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2018.08.010.

Malerba, D. (2008). A relational perspective on spatial data mining. *International Journal of Data Mining, Modelling and Management*, 1(1):103–118. https://dx.doi.org/10.1504/IJDMMM.2008.022540.

Manjula, A. and Narsimha, G. (2014). A review on spatial data mining methods and applications. *International Journal of Computer Engineering and Applications*, 7(I).

Martins, S. (2020). *Modelo de Proceso para Proyectos de Explotación de Información*. PhD thesis, Facultad de Informática, Universidad Nacional de La Plata.

Martins, S., Pesado, P., and García-Martínez, R. (2016). Information mining projects management process. In *International Conference on Software Engineering & Knowledge Engineering*, pages 504–509. https://doi.org/10.18293/SEKE2016-009.

McCann, C. (2015). Scaling airbnb with brian chesky — class 18 notes of stanford university's cs183c. https://medium.com/cs183c-blitzscaling-class-collection/scaling-airbnb-with-brian-chesky-class-18-notes-of-stanford-university-s-cs183c-3fcf75778358.

Menger, K. (1942). Statistical metrics. *Proceedings of the National Academy of Sciences of the United States of America*, 28(12):535.

Mennis, J. and Guo, D. (2009). Spatial data mining and geographic knowledge discovery—an introduction. *Computers, Environment and Urban Systems*, 33(6):403–408. https://doi.org/10.1016/j.compenvurbsys.2009.11.001.

Miller, H. J. and Han, J. (2009). *Geographic data mining and knowledge discovery*. CRC press.

Miyahara, S. and Miyamoto, S. (2014). A family of algorithms using spectral clustering and dbscan. In *2014 IEEE International Conference on Granular Computing (GrC)*, pages 196–200. IEEE. https://doi.org/10.1109/GRC.2014.6982834.

Needed, M. (2020). Airbnb by the numbers: Usage, demographics, and revenue growth. https://muchneeded.com/airbnb-statistics/. Información actualizada el 26 de mayo de 2020.

Ng, R. T. and Han, J. (1994). E cient and e ective clustering methods for spatial data mining. In *Proceedings of VLDB*, pages 144–155.

Ord, J. K. and Getis, A. (1995). Local spatial autocorrelation statistics: distributional issues and an application. *Geographical analysis*, 27(4):286–306.

Oxoli, D., Prestifilippo, G., Zurbaràn, M., and Shaji, S. (2019). Hotspot analysis plugin for qgis. Version 1.0.3. https://github.com/danioxoli/HotSpotAnalysis_Plugin.

Pastor-Satorras, R. and Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200. https://doi.org/10.1103/PhysRevLett.86.3200.

Pollo Cattaneo, M. F., Amatriain, H. G., Rodriguez, D., Pytel, P., Ciccolella, E., Vegega, C., Dearriba, M., Rodríguez Aubert, M., Bose, M., Giordano, F., et al. (2010). Ingeniería de proyectos de explotación de la información. In *XII Workshop de Investigadores en Ciencias de la Computación*.

Pyle, D. (2003). Business modeling and business intelligence. Morgan Kaufmann Publishers San Francisco.

Pytel, P., Britos, P., and García-Martínez, R. (2013). Proposal and validation of a feasibility model for information mining projects (s). In *International Conference on Software Engineering & Knowledge Engineering*, pages 83–88.

Pytel, P., Hossian, A., Britos, P., and García-Martínez, R. (2015). Feasibility and effort estimation models for medium and small size information mining projects. *Information Systems*, 47:1–14. https://doi.org/10.1016/j.is.2014.06.004.

Qgis (2020). Qgis. http://www.qgis.org.

Raifer, M. (2020). Overpass turbo. https://overpass-turbo.eu/.

RapidMiner, I. (2020). Rapidminer studio. http://www.rapidminer.com.

Rosen, K. H. and Krithivasan, K. (2012). *Discrete mathematics and its applications: with combinatorics and graph theory*. Tata McGraw-Hill Education, 7 edition.

Rottoli, G. (2020a). Airbnb dataset for spatial cluster detection. Mendeley Data, V1, doi: https://doi.org/10.17632/v4rvvz2ksf.1.

Rottoli, G. (2020b). Concepción del uruguay: Dataset for spatial association discovery. Mendeley Data, V1, https://doi.org/10.17632/jmct6vgvz8.1.

Rottoli, G. (2020c). San francisco: Dataset for co-location detection. Mendeley Data, V1, https://doi.org/10.17632/45bwn5tn7r.1.

Rottoli, G. (2020d). Us counties dataset for spatial outlier detection. Mendeley Data, V1, https://doi.org/10.17632/9vbgmp5swg.1.

Róttoli, G., Merlino, H., and García Martínez, R. (2018). Proceso para el descubrimiento de asociaciones espaciales mediante subgrafos frecuentes. In *XXIV Congreso Argentino de Ciencias de la Computación (La Plata, 2018).*

Rottoli, G. D. (2020e). Resúmenes técnicos de validación mediante juicios de expertos de procesos de explotación de información espacialmente referenciada. https://www.researchgate.net/publication/346611919_Resumenes_tecnicos_de_validacion_mediante_juicios_de_expertos_de_procesos_de_explotacion_de_informacion_espacialmente_referenciada.

Rottoli, G. D., Merlino, H., and García-Martínez, R. (2018). Knowledge discovery process for detection of spatial outliers. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 57–68. Springer. https://doi.org/10.1007/978-3-319-92058-0_6.

Rud, O. P. (2009). *Business intelligence success factors: tools for aligning your business in the global economy*, volume 18. John Wiley & Sons.

Sadat, Y. K., Nikaein, T., and Karimipour, F. (2015). Fuzzy spatial association rule mining to analyze the effect of environmental variables on the risk of allergic asthma prevalence. *Geodesy and Cartography*, 41(2):101–112. https://doi.org/10.3846/20296991.2015.1051339.

Sagvekar, V., Sagvekar, V., and Deorukhkar, K. (2013). Performance assessment of clarans: A method for clustering objects for spatial data mining. *Global journal of engineering, design and technology*, 2(6):1–8.

Schröer, C., Kruse, F., and Gómez, J. M. (2021). A systematic literature review on applying crisp-dm process model. *Procedia Computer Science*, 181:526–534.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., and Xu, X. (2017). Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21. https://doi.org/10.1145/3068335.

Schubert, E., Zimek, A., and Kriegel, H.-P. (2014). Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data Mining and Knowledge Discovery*, 28(1):190–237. https://doi.org/10.1007/s10618-012-0300-z.

Sha, Z. and Li, X. (2010). Mining local association patterns from spatial dataset. In *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, volume 3, pages 1455–1460. IEEE. https://doi.org/10.1109/FSKD.2010.5569205.

Shekhar, S., Evans, M. R., Kang, J. M., and Mohan, P. (2011). Identifying patterns in spatial information: A survey of methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):193–214. https://doi.org/10.1002/widm.25.

Shekhar, S. and Huang, Y. (2001). Discovering spatial co-location patterns: A summary of results. In *International symposium on spatial and temporal databases*, pages 236–256. Springer. https://doi.org/10.1007/3-540-47724-1_13.

Shi, T. (2019). Spatial data mining and big data analysis of tourist travel behavior. *Ingénierie des Systèmes d'Information*, 24(2):167–172. https://doi.org/10.18280/isi.240206.

Shi, Y., Deng, M., Yang, X., and Liu, Q. (2016). Adaptive detection of spatial point event outliers using multilevel constrained delaunay triangulation. *Computers, Environment and Urban Systems*, 59:164–183.

Shi, Y., Gong, J., Deng, M., Yang, X., and Xu, F. (2018). A graph-based approach for detecting spatial cross-outliers from two types of spatial point events. *Computers, Environment and Urban Systems*, 72:88–103.

Singh, A. K. and Lalitha, S. (2018). A novel spatial outlier detection technique. *Communications in Statistics-Theory and Methods*, 47(1):247–257.

Song, H. and Lee, J.-G. (2018). Rp-dbscan: A superfast parallel dbscan algorithm based on random partitioning. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1173–1187. https://doi.org/10.1145/3183713.3196887.

Strash, D. (2018). Quick cliques: Quickly compute all maximal cliques in sparse graphs. https://github.com/darrenstrash/quick-cliques.

Sutjipto, S. S. U., Sitanggang, I. S., and Barus, B. (2017). Potential usage estimation of ground water using spatial association rule mining. *Telkomnika*, 15(1):504.

Tang, J., Chen, Z., Fu, A. W.-C., and Cheung, D. W. (2002). Enhancing effectiveness of outlier detections for low density patterns. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 535–548. Springer. https://doi.org/10.1007/3-540-47887-6_53.

TarakaSravani, B., Srineha, S., Tejaswini, G., Raj, K. K., Santhi, M., Vaddeswaram, G., and Pradesh, A. (2017). Spatial data mining using clustering techniques. *International Journal of Pure and Applied Mathematics*, 115(8):425–430.

Tobler, W. R. (1979). Cellular geography. In *Philosophy in geography*, pages 379–386. Springer. https://doi.org/10.1007/978-94-009-9394-5_18.

Tripathy, M. and Panda, A. (2017). A study of algorithm selection in data mining using meta-learning. *Journal of Engineering Science & Technology Review*, 10(2).

Tung, A. K., Hou, J., and Han, J. (2001). Spatial clustering in the presence of obstacles. In *Proceedings 17th International Conference on Data Engineering*, pages 359–367. IEEE.

Urquiza Almandor, O. F. (2002a). *Historia de Concepción del Uruguay: 1783 - 1826*, volume I. Biblioteca de la Respetable Logia Jorge Washington N 44, first edition. Comisión Técnica Mixta de Salto Grande, Delegación Argentina.

Urquiza Almandor, O. F. (2002b). *Historia de Concepción del Uruguay: 1826 - 1870*, volume II. Biblioteca de la Respetable Logia Jorge Washington N 44, first edition. Comisión Técnica Mixta de Salto Grande, Delegación Argentina.

Urquiza Almandor, O. F. (2002c). *Historia de Concepción del Uruguay: 1871 - 1890*, volume III. Biblioteca de la Respetable Logia Jorge Washington N 44, first edition. Comisión Técnica Mixta de Salto Grande, Delegación Argentina.

US Census Bureau (2019). Cartographic boundary shapefiles. https://www.census.gov/geographies/mapping-files/2016/geo/carto-boundary-file.html.

U.S. Census Bureau (2020a). Annual estimates of the resident population for counties: April 1, 2010 to july 1, 2013. En inglés. http://www.census.gov/popest/data/counties/totals/2013/CO-EST2013-01.html (30 de octubre de 2020).

U.S. Census Bureau (2020b). Quickfacts: United states. united states census. En inglés. https://www.census.gov/quickfacts/fact/table/US/PST045219 (14 de agosto de 2020).

Van Zoest, V., Stein, A., and Hoek, G. (2018). Outlier detection in urban air quality sensor networks. *Water, Air, & Soil Pollution*, 229(4):111. https://doi.org/10.1007/s11270-018-3756-7.

Varghese, B. M., Unnikrishnan, A., and Jacob, K. (2013). Spatial clustering algorithms-an overview. *Asian Journal of Computer Science and Information Technology*, 3(1):1–8.

Vijayalaksmi, S. and Punithavalli, M. (2012). A fast approach to clustering datasets using dbscan and pruning algorithms. *International Journal of Computer Applications*, 60(14).

Wang, J. and Wang, X. (2012). A spatial clustering method for points-with-directions. In *International Conference on Rough Sets and Knowledge Technology*, pages 194–199. Springer.

Wang, L., Bao, Y., Lu, J., and Yip, J. (2008). A new join-less approach for co-location pattern mining. In *2008 8th IEEE International Conference on Computer and Information Technology*, pages 197–202. IEEE.

Wang, W., Tang, M., Stanley, H. E., and Braunstein, L. A. (2017). Unification of theoretical approaches for epidemic spreading on complex networks. *Reports on Progress in Physics*, 80(3):036603. https://doi.org/10.1088/1361-6633/aa5398.

Washio, T. and Motoda, H. (2003). State of the art of graph-based data mining. *Acm Sigkdd Explorations Newsletter*, 5(1):59–68. https://doi.org/10.1145/959242.959249.

Wei, G., Liu, H., and Xie, M. (2009). Clustering large spatial data with local-density and its application. *Information Technology Journal*, 8(4):476–485.

Wieringa, R. J. (2014). *Design science methodology for information systems and software engineering*. Springer. https://doi.org/10.1007/978-3-662-43839-8.

Xu, W., Gao, H., Liu, Y., and Li, L. (2017). An adaptive spatial outlier detection algorithm with no parameter for wsn. In *2017 20th International Conference on Information Fusion (Fusion)*, pages 1–8. IEEE.

Xu, Z. (2008). *Linguistic Aggregation Operators: An Overview*, pages 163–181". Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-73723-0_9.

Yager, R. R. (1988). On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on systems, Man, and Cybernetics*, 18(1):183–190.

Yang, H., Parthasarathy, S., and Mehta, S. (2005). Mining spatial object associations for scientific data. In *IJCAI*, pages 902–907.

Yoo, J. S., Shekhar, S., and Celik, M. (2005). A join-less approach for co-location pattern mining: A summary of results. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 4–pp. IEEE.

Yu, W. (2016). Spatial co-location pattern mining for location-based services in road networks. *Expert Systems with Applications*, 46:324–335.

Zeitouni, K. (2002). A survey of spatial data mining methods databases and statistics point of views. In *Data warehousing and web engineering*, pages 229–242. IGI Global. https://doi.org/10.4018/978-1-931777-02-5.ch013.

Zhang, B., Yin, W. J., Xie, M., and Dong, J. (2007). Geo-spatial clustering with non-spatial attributes and geographic non-overlapping constraint: a penalized spatial distance measure. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 1072–1079. Springer.

Zheng, G., Brantley, S. L., Lauvaux, T., and Li, Z. (2017). Contextual spatial outlier detection with metric learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2161–2170.

APÉNDICE A

PREGUNTAS GUÍAS

Debido a que el proceso de recolección de datos ha sido por medio de entrevistas no estructuradas, las preguntas que a continuación se detallan no han sido realizadas de forma rigurosa. Mas bien, por el contrario, estas cuestiones han servido de guía o *checklist* para dirigir la atención del experto entrevistado hacia aspectos determinados del artefacto evaluado. Estas preguntas poseen la intención de que el experto pueda considerar todos los ángulos para la asignación de un valor cualitativo difuso a cada uno de los atributos evaluados. Las mismas fueron expresamente utilizadas en caso de haber sido necesario si el entrevistador no dio cuenta de tal información en las respuestas abiertas del entrevistado.

1. En cuanto a la estructura, la comprensibilidad de la misma y la utilidad de los procesos:

   a) ¿La estructura del proceso resulta comprensible?

   b) ¿Cree que los procesos se encuentran suficientemente detallados?

   c) ¿Cree que es útil esta secuencia de pasos?

   d) ¿Cada actividad tiene un sentido determinado para usted?

   e) ¿Cree que hace falta alguna actividad que de cuenta de algún aspecto que haya sido ignorado?

   f) ¿Cree que sobra alguna actividad o resulta redundante?

   g) ¿Considera que algún aspecto relacionado con los datos espaciales ha sido ignorado?

2. En cuanto a la comprensibilidad e interpretabilidad de los patrones:

   a) ¿Comprende cada una de las salidas de las actividades del proceso?

   b) ¿Le es posible derivar conclusiones a partir de ellas?

   c) ¿Cree que es lógica la forma de representar la información??

   d) ¿Cómo cree que se pueden combinar las salidas parciales en la etapa de evaluación? ¿Le parece intuitivo?

   e) ¿Considera que son necesarios algunos cambios en cuanto a las salidas? ¿Cuáles?

   f) ¿Considera que algún aspecto relacionado con la información espacial ha sido ignorado?

3. En cuanto a la adaptabilidad de los procesos:

   a) ¿Conoce herramientas que puedan ser útiles para cada actividad?

   b) ¿Cree que son suficientemente versátiles?

   c) ¿Cree que sería difícil de implementar alguna de estas actividades?

   d) ¿Qué tan complicado considera que es integrar varias herramientas para implementar el proceso?

   e) ¿Cree que alguna herramienta necesaria sería difícil de adquirir o integrar?

   f) ¿Considera que puede implementarse completamente mediante código fuente en algún lenguaje de programación?

   g) ¿Considera valiosa la modularidad del diseño?

   h) ¿Cómo cree que reaccionará el proceso ante distintos tipos de datos espaciales?

   i) ¿Cómo cree que reaccionará el proceso ante distintos tipos de relaciones espaciales?

   j) ¿Cómo cree que reaccionará el proceso ante cambios en la definición de los vecindarios?

APÉNDICE B

CÁLCULOS AUXILIARES

Los siguientes cálculos auxiliares se corresponden con lo expuesto en la sección 5.5.4. El fundamento teórico de los mismos se encuentra especificado en la sección 5.5.1.

B.1. Ordenamiento de las dimensiones evaluadas

$$
\begin{aligned}
S &= (s_0 : \textit{Muy Bajo}, s_1 : \textit{Bajo}, s_2 : \textit{Medio}, s_3 : \textit{Alto}, s_4 : \textit{Muy Alto}) \\
d &= (\textit{Utilidad}, \textit{Interpretabilidad}, \textit{Adaptabilidad}) \\
I(d_i) &= \{(d_1, \textit{Muy Alto}), (d_2, \textit{Medio}), (d_3, \textit{Alto})\} \\
\alpha &= 1
\end{aligned}
$$

$$
\begin{aligned}
Sup_1 &= Sup(\textit{Muy Alto}, \textit{Muy Alto}) + Sup(\textit{Muy Alto}, \textit{Medio}) + Sup(\textit{Muy Alto}, \textit{Alto}) \\
Sup_1 &= 1 + 0 + 1 \\
Sup_1 &= 2 \\
Sup_2 &= Sup(\textit{Medio}, \textit{Muy Alto}) + Sup(\textit{Medio}, \textit{Medio}) + Sup(\textit{Medio}, \textit{Alto}) \\
Sup_2 &= 0 + 1 + 1 \\
Sup_2 &= 2 \\
Sup_3 &= Sup(\textit{Alto}, \textit{Muy Alto}) + Sup(\textit{Alto}, \textit{Medio}) + Sup(\textit{Alto}, \textit{Alto}) \\
Sup_3 &= 1 + 1 + 1 \\
Sup_3 &= 3
\end{aligned}
$$

$$
\begin{aligned}
u_1(I(d_1)) &= [Sup_1 + Ind(I(d_1))]/2 = [2+4]/2 = 3 \\
u_2(I(d_1)) &= [Sup_2 + Ind(I(d_2))]/2[2+2]/2 = 2 \\
u_3(I(d_3)) &= [Sup_3 + Ind(I(d_3))]/2 = [3+3]/2 = 3
\end{aligned}
$$

$$
\sigma(i) \quad : \quad \{(2,1),(1,2),(3,3)\}
$$

$$
w_1 = \frac{Q\left(\frac{2}{3}\right)}{Q\left(\frac{2}{3}\right) + Q\left(\frac{1}{1}\right) + Q\left(\frac{1}{1}\right)} = \frac{0.92}{0.92+1+1} = 0.315
$$

$$
w_2 = \frac{Q\left(\frac{1}{1}\right)}{Q\left(\frac{1}{1}\right) + Q\left(\frac{1}{1}\right) + Q\left(\frac{1}{1}\right)} = \frac{1}{0.92+1+1} = 0.342
$$

$$
w_3 = \frac{Q\left(\frac{1}{1}\right)}{Q\left(\frac{1}{1}\right) + Q\left(\frac{1}{1}\right) + Q\left(\frac{1}{1}\right)} = \frac{1}{0.92+1+1} = 0.342
$$

$$
W = (0.32, 0.34, 0.34)
$$

B.2. Agregación de dimensiones evaluadas por experto

$$
\begin{aligned}
r_1^1 &= S(Redondear(0.32 * Ind(\textbf{\textit{Muy Alto}}) + 0.34 * Ind(\textbf{\textit{Medio}}) + 0.34 * Ind(\textbf{\textit{Alto}}))) = \\
&= S(Redondear(0.32 * 4 + 0.34 * 2 + 0.34 * 3)) = \\
&= S(Redondear(1.28 + 0.68 + 1.02)) = S(3) = \\
&= \textbf{\textit{Alto}}
\end{aligned}
$$

$$r_2^1 = S(Redondear(0.32 * Ind(\textbf{\textit{Muy Alto}}) + 0.34 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Muy Alto}}))) =$$
$$= S(Redondear(0.32 * 4 + 0.34 * 3 + 0.34 * 4)) =$$
$$= S(Redondear(1.28 + 1.02 + 1.36)) = S(4) =$$
$$= \textbf{\textit{Muy Alto}}$$
$$r_3^1 = S(Redondear(0.32 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Medio}}) + 0.34 * Ind(\textbf{\textit{Alto}}))) =$$
$$= S(Redondear(0.32 * 3 + 0.34 * 2 + 0.34 * 3)) =$$
$$= S(Redondear(0.96 + 0.68 + 1.02)) = S(3) =$$
$$= \textbf{\textit{Alto}}$$
$$r_4^1 = S(Redondear(0.32 * Ind(\textbf{\textit{Muy Alto}}) + 0.34 * Ind(\textbf{\textit{Bajo}}) + 0.34 * Ind(\textbf{\textit{Alto}}))) =$$
$$= S(Redondear(0.32 * 4 + 0.34 * 1 + 0.34 * 3)) =$$
$$= S(Redondear(1.28 + 0.34 + 1.02)) = S(3) =$$
$$= \textbf{\textit{Alto}}$$
$$r_5^1 = S(Redondear(0.32 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Alto}}))) =$$
$$= S(Redondear(0.32 * 3 + 0.34 * 3 + 0.34 * 3)) =$$
$$= S(Redondear(0.96 + 1.02 + 1.02)) = S(3) =$$
$$= \textbf{\textit{Alto}}$$
$$r_1^2 = S(Redondear(0.32 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Alto}}))) =$$
$$= S(Redondear(0.32 * 3 + 0.34 * 3 + 0.34 * 3)) =$$
$$= S(Redondear(0.96 + 1.02 + 1.02)) = S(3) =$$
$$= \textbf{\textit{Alto}}$$
$$r_2^2 = S(Redondear(0.32 * Ind(\textbf{\textit{Medio}}) + 0.34 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Alto}}))) =$$
$$= S(Redondear(0.32 * 2 + 0.34 * 3 + 0.34 * 3)) =$$
$$= S(Redondear(0.64 + 1.02 + 1.02)) = S(3) =$$
$$= \textbf{\textit{Alto}}$$
$$r_3^2 = S(Redondear(0.32 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Muy Alto}}) + 0.34 * Ind(\textbf{\textit{Muy Alto}}))) =$$
$$= S(Redondear(0.32 * 3 + 0.34 * 4 + 0.34 * 4)) =$$
$$= S(Redondear(0.96 + 1.36 + 1.36)) = S(4) =$$
$$= \textbf{\textit{Muy Alto}}$$

$$r_4^2 = S(Redondear(0.32 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Muy Alto}}) + 0.34 * Ind(\textbf{\textit{Alto}}))) =$$
$$= S(Redondear(0.32 * 3 + 0.34 * 4 + 0.34 * 3)) =$$
$$= S(Redondear(0.96 + 1.36 + 1.02)) = S(3) =$$
$$= \textbf{\textit{Alto}}$$
$$r_5^2 = S(Redondear(0.32 * Ind(\textbf{\textit{Medio}}) + 0.34 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Alto}}))) =$$
$$= S(Redondear(0.32 * 2 + 0.34 * 3 + 0.34 * 3)) =$$
$$= S(Redondear(0.64 + 1.02 + 1.02)) = S(3) =$$
$$= \textbf{\textit{Alto}}$$
$$r_1^3 = S(Redondear(0.32 * Ind(\textbf{\textit{Medio}}) + 0.34 * Ind(\textbf{\textit{Muy Alto}}) + 0.34 * Ind(\textbf{\textit{Bajo}}))) =$$
$$= S(Redondear(0.32 * 2 + 0.34 * 4 + 0.34 * 1)) =$$
$$= S(Redondear(0.64 + 1.36 + 0.34)) = S(2) =$$
$$= \textbf{\textit{Medio}}$$
$$r_2^3 = S(Redondear(0.32 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Muy Alto}}) + 0.34 * Ind(\textbf{\textit{Medio}}))) =$$
$$= S(Redondear(0.32 * 3 + 0.34 * 4 + 0.34 * 2)) =$$
$$= S(Redondear(0.96 + 1.36 + 0.68)) = S(3) =$$
$$= \textbf{\textit{Alto}}$$
$$r_3^3 = S(Redondear(0.32 * Ind(\textbf{\textit{Medio}}) + 0.34 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Alto}}))) =$$
$$= S(Redondear(0.32 * 2 + 0.34 * 3 + 0.34 * 3)) =$$
$$= S(Redondear(0.64 + 1.02 + 1.02)) = S(3) =$$
$$= \textbf{\textit{Alto}}$$
$$r_4^3 = S(Redondear(0.32 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Medio}}))) =$$
$$= S(Redondear(0.32 * 3 + 0.34 * 3 + 0.34 * 2)) =$$
$$= S(Redondear(0.96 + 1.02 + 0.68)) = S(3) =$$
$$= \textbf{\textit{Alto}}$$
$$r_5^3 = S(Redondear(0.32 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Muy Alto}}) + 0.34 * Ind(\textbf{\textit{Alto}}))) =$$
$$= S(Redondear(0.32 * 3 + 0.34 * 4 + 0.34 * 3)) =$$
$$= S(Redondear(0.96 + 1.36 + 1.02)) = S(3) =$$
$$= \textbf{\textit{Alto}}$$

$$r_1^4 = S(Redondear(0.32 * Ind(\textbf{\textit{Medio}}) + 0.34 * Ind(\textbf{\textit{Muy Alto}}) + 0.34 * Ind(\textbf{\textit{Bajo}}))) =$$
$$= S(Redondear(0.32 * 2 + 0.34 * 4 + 0.34 * 1)) =$$
$$= S(Redondear(0.64 + 1.36 + 0.34)) = S(2) =$$
$$= \textbf{\textit{Medio}}$$
$$r_2^4 = S(Redondear(0.32 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Medio}}))) =$$
$$= S(Redondear(0.32 * 3 + 0.34 * 3 + 0.34 * 2)) =$$
$$= S(Redondear(0.96 + 1.02 + 0.68)) = S(3) =$$
$$= \textbf{\textit{Alto}}$$
$$r_3^4 = S(Redondear(0.32 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Medio}}))) =$$
$$= S(Redondear(0.32 * 3 + 0.34 * 3 + 0.34 * 2)) =$$
$$= S(Redondear(0.96 + 1.02 + 0.68)) = S(3) =$$
$$= \textbf{\textit{Alto}}$$
$$r_4^4 = S(Redondear(0.32 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Medio}}))) =$$
$$= S(Redondear(0.32 * 3 + 0.34 * 3 + 0.34 * 2)) =$$
$$= S(Redondear(0.96 + 1.02 + 0.68)) = S(3) =$$
$$= \textbf{\textit{Alto}}$$
$$r_5^4 = S(Redondear(0.32 * Ind(\textbf{\textit{Alto}}) + 0.34 * Ind(\textbf{\textit{Muy Alto}}) + 0.34 * Ind(\textbf{\textit{Alto}}))) =$$
$$= S(Redondear(0.32 * 3 + 0.34 * 4 + 0.34 * 3)) =$$
$$= S(Redondear(0.96 + 1.36 + 1.02)) = S(3) =$$
$$= \textbf{\textit{Alto}}$$

B.3. Agregación de expertos por proceso

B.3.1. Proceso 1

$$P_1 = (\textit{Alto}, \textit{Muy Alto}, \textit{Alto}, \textit{Alto}, \textit{Alto})$$
$$u_i = (Sup(\textit{Alto}), Sup(\textit{Muy Alto}), Sup(\textit{Alto}), Sup(\textit{Alto}), Sup(\textit{Alto})) =$$
$$= (5, 5, 5, 5, 5)$$
$$\sigma : \{(i, i)\}, \forall i = 1..5$$
$$w_i = \frac{Q\left(\frac{5}{5}\right)}{5 * Q\left(\frac{5}{5}\right)} = \frac{1}{5} = 0.2, \forall i = 1..5$$
$$W = (0.2, 0.2, 0.2, 0.2, 0.2)$$

B.3.2. Proceso 2

$$P_2 = (\textit{Alto}, \textit{Alto}, \textit{Muy Alto}, \textit{Alto}, \textit{Alto})$$
$$u_i = (Sup(\textit{Alto}), Sup(\textit{Alto}), Sup(\textit{ Muy Alto}), Sup(\textit{Alto}), Sup(\textit{Alto})) =$$
$$= (5, 5, 5, 5, 5)$$
$$\sigma : \{(i, i)\}, \forall i = 1..5$$
$$w_i = \frac{Q\left(\frac{5}{5}\right)}{5 * Q\left(\frac{5}{5}\right)} = \frac{1}{5} = 0.2, \forall i = 1..5$$
$$W = (0.2, 0.2, 0.2, 0.2, 0.2)$$

B.3.3. Procesos 3 y 4

$$
\begin{aligned}
P_3 = P_4 &= (\textit{Medio}, \textit{Alto}, \textit{Alto}, \textit{Alto}, \textit{Alto}) \\
u_i &= (Sup(\textit{Alto}), Sup(\textit{Muy Alto}), Sup(\textit{Alto}), Sup(\textit{Alto}), Sup(\textit{Alto})) = \\
&= (5, 5, 5, 5, 5) \\
\sigma &: \{(i, i)\}, \forall i = 1..5 \\
w_i &= \frac{Q\left(\frac{5}{5}\right)}{5 * Q\left(\frac{5}{5}\right)} = \frac{1}{5} = 0.2, \forall i = 1..5 \\
W &= (0.2, 0.2, 0.2, 0.2, 0.2)
\end{aligned}
$$

B.3.4. Ponderación

$$
\begin{aligned}
P^1 &= S(Redondear(0.2 * [3 + 4 + 3 + 3 + 3])) = S(Redondear(0.2 * 16)) = \\
&= S(Redondear(3.2)) = \textit{Alto} \\
P^2 &= S(Redondear(0.2 * [3 + 3 + 4 + 3 + 3])) = S(Redondear(0.2 * 16)) = \\
&= S(Redondear(3.2)) = \textit{Alto} \\
P^3 &= S(Redondear(0.2 * [2 + 3 + 3 + 3 + 3])) = S(Redondear(0.2 * 14)) = \\
&= S(Redondear(2.8)) = \textit{Alto} \\
P^4 &= S(Redondear(0.2 * [2 + 3 + 3 + 3 + 3])) = S(Redondear(0.2 * 14)) = \\
&= S(Redondear(2.8)) = \textit{Alto}
\end{aligned}
$$

APÉNDICE C

RECURSOS EXTERNOS

C.1. Bases de datos

Las siguientes bases de datos fueron utilizadas para la demostración de los artefactos propuestos.

- "AirBnB Dataset for Spatial Cluster Detection"
  Mendeley Data, V1, doi: 10.17632/v4rvvz2ksf.1
  http://dx.doi.org/10.17632/v4rvvz2ksf.1

- "San Francisco Dataset for Co-location Detection"
  Mendeley Data, V1, doi: 10.17632/45bwn5tn7r.1
  http://dx.doi.org/10.17632/45bwn5tn7r.1

- "Concepción del Uruguay Dataset for Spatial Association Discovery"
  Mendeley Data, V1, doi: 10.17632/jmct6vgvz8.1
  http://dx.doi.org/10.17632/jmct6vgvz8.1

- "US Counties dataset for Spatial Outlier Detection "
  Mendeley Data, V1, doi: 10.17632/9vbgmp5swg.1
  http://dx.doi.org/10.17632/9vbgmp5swg.1

C.2. Reportes

Los reportes técnicos producto de la validación por juicio de expertos se encuentran disponibles en el siguiente repositorio:

- Rottoli, Giovanni Daián (2020). Resúmenes técnicos de validación mediante juicios de expertos de procesos de explotación de información espacialmente referenciada. Reporte Técnico. Online: `https://www.researchgate.net/publication/346611919_Resumenes_tecnicos_de_validacion_mediante_juicios_de_expertos_de_procesos_de_explotacion_de_informacion_espacialmente_referenciada`

APÉNDICE D

HERRAMIENTAS SOFTWARE

Las herramientas presentadas en este anexo son factibles de ser utilizadas en las distintas actividades de los procesos de explotación de información propuestos en la tesis doctoral. De todas formas, resulta necesario realizar un análisis de factibilidad para determinar si las mismas satisfacen los requerimientos según el problema de dominio abordado y seleccionar la más adecuada a tales efectos.

La lista presentada a continuación tiene como propósito ejemplificar la variedad de opciones, mas no provee un compendio exhaustivo de todas las opciones del estado de la cuestión.

Los hipervínculos provistos por cada opción fueron accedidos el día 9 de septiembre de 2021 para corroborar su correcto funcionamiento, mas no se garantiza que eventualmente los mismos queden obsoletos.

D.1. Lenguajes de programación

- Python. `https://www.python.org/`.
- The R Project for Statistical Computing. `https://www.r-project.org/`.
- The Scala Programming Language. `https://www.scala-lang.org/`.

D.2. Herramientas de minería de datos

- Rapidminer Studio. https://rapidminer.com/
- Weka 3. https://www.cs.waikato.ac.nz/ml/weka/
- Knime. https://www.knime.com/
- Orange Data Mining https://orange.biolab.si/.
- Tanagra Project. http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html
- Pandas Lib. https://pandas.pydata.org/.
- Scikit-learn. Machine Learning in Python. https://scikit-learn.org.

D.3. Herramientas de manejo de información espacial

- Geoda Tool. https://geodacenter.github.io/.
- Qgis Project https://qgis.org/es/site/.
- ArcGIS Project https://www.arcgis.com/index.html.
- GeoPandas https://geopandas.org/.

D.4. Herramientas de manejo de grafos

- Subdue: Graph-based Unsupervised Learning (Discovery). http://ailab.wsu.edu/subdue/
- gSpan: Frequent Graph Mining Package. https://sites.cs.ucsb.edu/~xyan/software/gSpan.htm
- Quick Cliques: Quickly compute all maximal cliques in sparse graphs https://github.com/darrenstrash/quick-cliques
- Graph Mining tools. http://graphmining.ailab.wsu.edu/node/9/index.html

- Google Research Library for Graph Mining. `https://research.google/teams/algorithms-optimization/graph-mining/`
- JUNG: Java Universal Network/Graph Framework. `http://jung.sourceforge.net/`.
- MIPS: The migrant prototype system. `https://dtai.cs.kuleuven.be/software/mips`

“So, like everyone else, I was staring out of one of the windows of the inn at the end of the **words**. Worlds. I meant worlds.”

—Brant Tucker, in Sandman 56: “World’s End”