UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS
Facultad 3

Método de estratificación de territorios basado en Sistemas de Información Geográfica y medidas de similitud geométrica

Trabajo final presentado en opción al título de

Máster en Informática Avanzada

Autor: Ing. Liset González Polanco

Tutor: Dra.C. Roxana Cañizarez González

La Habana, enero de 2019

Declaración de autoría

Declaro por este medio que yo Liset González Polanco, con carné de identidad 86072120052, soy la autora principal del trabajo final de maestría Método de estratificación de territorios basado en Sistemas de Información Geográfica y medidas de similitud geométrica, desarrollada como parte de la Maestría en Informática Avanzada y que autorizo a la Universidad de las Ciencias Informáticas a hacer uso de la misma en su beneficio, así como los derechos patrimoniales con carácter exclusivo.

Y para que así conste, firmo la presente declaración jurada de autoría en La Habana a los 31 días del mes de enero del año 2019

Ing. Liset González Polanco _____

Dra.C. Roxana Cañizarez González _____

AGRADECIMIENTOS

A la Revolución Cubana por llenar de oportunidades a los jóvenes cubanos.

A Fidel Castro mi eterno comandante que con su visión única y futurista fundó la universidad de las Ciencias Informáticas, casa de altos estudios donde me he formado, con un claustro comprometido y revolucionario, del cual que me siento orgullosa

A mis padres, hermana,tías, primas, abuelas y abuelos que fueron mi primera escuela, puerto seguro, para siempre dar gracias a la vida como Violeta Parra en la voz de Sonita.

A Yadian por entrelazar conmigo sus manos para seguir juntos en esta y otras aspiraciones, amparo en mis desvelos, sosiego en cada dilema.

A mi pequeña querida, faro de luz de cada mañana.

A mi tutora, mentora y amiga, gracias por brindarme su guía.

A todos los que de una forma u otra han sido partícipes de este trabajo.

DEDICATORIA

A mi idolatrada Liety.

RESUMEN

Una meta del sistema de salud es la prevención de enfermedades, por ello cobra especial importancia el estudio de la relación de enfermedades con el espacio. Existen evidencias del empleo de los Sistemas de Información Geográfica en estudios sobre la distribución espacial de problemas de salud. A pesar de esto, los trabajos reportados en la literatura consultada no explotan la componente espacial de los datos, lo que viola el principio de la primera ley de la geografía. Por otra parte, existe dispersión en las metodologías, herramientas y técnicas para abordar estudios de este tipo.

En esta investigación se presenta un método de estratificación de territorios basado en Sistemas de Información Geográfica y medidas de similitud geométrica, definidas a partir de los criterios: distancia, tamaño y conectividad. La propuesta permite realizar estudios estratificados según la primera ley de la geografía y garantiza la obtención de estratos más compactos. El método propuesto cuenta con cinco etapas: Selección de indicadores y territorios, Preprocesamiento de indicadores, Agrupamiento, Postprocesamiento y Visualización, soportado en una solución informática basada en software libre. Como parte de la validación se aplica el método en dos casos de estudios y se realiza el análisis de índices que avalan la efectividad de la propuesta.

Palabras claves: estratificación de territorios, medidas de similitud geométricas, Sistema de Información Geográfica, técnicas de agrupamiento, problemas de salud.

ABSTRACT

A goal of the health system is the prevention of diseases, which is why the study of the relationship of diseases with space is particularly important. There is evidence of the use of Geographic Information Systems in studies on the spatial distribution of health problems. Despite this, the works reported in the consulted literature do not exploit the spatial component of the data, which violates the principle of the first law of geography. On the other hand, there is dispersion in the methodologies, tools and techniques to approach studies of this type.

This research presents a method of stratification of territories based on Geographic Information Systems and measures of geometric similarity, defined from the criteria: distance, size and connectivity. The proposal allows stratified studies according to the first law of geography and guarantees the obtaining of more compact strata. The proposed method has five stages: Selection, Preprocessing, Grouping, Postprocessing and Visualization, supported in a software solution based on free software. As part of the validation the method is applied in two cases of studies and the analysis of indexes that guarantee the effectiveness of the proposal is carried out.

Key Words: stratification of territories, geometric similarity measures, Geographic Information Systems, clustering techniques, health risks.

Índice general

Índice de Figuras

INTRODUCCIÓN

Históricamente se ha podido establecer la relación entre el espacio y los problemas de salud. A nivel mundial se observan enfermos, pero en algunas regiones se registran casos con más frecuencia que en otras. Un ejemplo de la relación del espacio y las enfermedades lo constituye el estudio realizado por Jhon Snow sobre el cólera en 1854 [4, 5] en Londres, donde establece una relación entre los fallecimientos ocurridos y los suministros de agua.

El conocimiento de la distribución geográfica permite establecer políticas de salud en la atención de una enfermedad y las prioridades necesarias en cuanto a recursos [6, 7, 8]. En este sentido, la estratificación de territorio denota como una valiosa herramienta para analizar el comportamiento de variables en el espacio. Es considerada como un procedimiento que permite clasificar objetos en clases homogéneas a partir de analogías o relaciones que se establecen entre sus características [9, 10, 11, 12, 13, 14, 15, 16]. En estudios salubristas suele denominarse estratificación epidemiológica y es parte del proceso integrado de diagnóstico-intervención-evaluación [17, 18, 3, 19, 20].

La estratificación epidemiológica utiliza varios enfoques para contribuir a la selección de sitios o zonas con problemas de salud y planificar estrategias de intervención [21, 22]. Dentro de los enfoques utilizados se encuentra la estratificación del riesgo [23, 24] y la del riesgo absoluto [23], por indicadores ponderados, así como los patrones de distribución de frecuencia de los principales riesgos asociados y las técnicas de análisis multivariado basados en análisis de grupos o agrupamiento [25, 26].

La utilización de indicadores ponderados ha sido ampliamente utilizada en estudios salubristas [14, 20]. Este método utiliza un conjunto de indicadores o factores de riesgos asociados al estudio. Los indicadores son ponderados con pesos obtenidos por criterio de expertos en el campo y se obtienen valores para cada territorio, que son utilizados para construir los grupos a partir de rangos establecidos. La principal limitación de esta estrategia radica en el sesgo que se puede introducir al definir los pesos y los rangos para construir los grupos.

Los patrones de distribución de frecuencia de los principales riesgos asociados se utilizan fundamentalmente cuando no se puede determinar o no están identificados los indicadores del riesgo. Su objetivo es identificar zonas donde determinado factor tiene una mayor frecuencia de aparición y luego acometer acciones de intervención [27].

Cuando es posible cuantificar los indicadores de riesgo y su influencia sobre el área se ha utilizado la estratificación epidemiológica del riesgo. Esta estrategia permite obtener una evaluación sobre el nivel de reducción del problema si se actúa sobre los factores de riesgo. Su principal limitación es que no permite identificar cambios sobre los

grupos pues constituye una fotografía del problema en estudio, además suele ser compleja debido a la necesidad de estudios para determinar el riesgo relativo y el riesgo atribuible poblacional [28].

La distribución del riesgo absoluto se basa en la utilización de un solo indicador para construir los estratos a partir de rangos de distribución según la tasa de incidencia o característica de la región de estudio. Es muy utilizada para la vigilancia epidemiológica a corto plazo pues permite monitorizar determinados indicadores en áreas priorizadas de forma rápida, práctica y oportuna [21, 29].

Las técnicas de análisis multivariado permiten realizar estudios más complejos porque pueden incorporarse numerosos factores de riesgos y otras variables asociadas al problema. Desde este enfoque la utilización de las técnicas de agrupamiento han sido ampliamente reportadas en la literatura [11, 30, 31]. Se debe especificar que este tratamiento no permite describir relaciones espaciales sobre objetos y por tanto dificulta la incorporación del espacio en el proceso.

En la literatura consultada se propone incorporar la naturaleza espacial a partir de una transformación sobre objetos puntuales o líneas, en la que luego son tratados como temáticos [3]. Esta propuesta no está alineada a la primera ley de la geografía donde se establece que los objetos en el espacio están relacionados, pero objetos cercanos están más relacionados entre sí que objetos distantes [32, 33]. Se aborda que el enfoque a utilizar en el proceso de estratificación depende de la urgencia de los resultados para tomar decisiones, la disponibilidad de información asociada al problema en estudio, los recursos disponibles, el entrenamiento de los especialistas que acometerán el estudio y su finalidad [34]. La utilización de los Sistemas de Información Geográfica (SIG) en el análisis de la distribución espacial de enfermedades ha aumentado considerablemente, sustentado en las herramientas de análisis existentes que posibilitan resolver problemas asociados a la distribución espacial [35].

Los SIG son herramientas básicas para la confección de mapas digitales y para los análisis geoespaciales, en todas las esferas del saber, que van más allá de análisis estadísticos y que tributan a una mejor planificación de infraestructura por ejemplo en: estudio demográfico, análisis de vías de transporte, distribución de recursos, distribución y comportamiento de enfermedades en salud [36].

El desarrollo de los SIG y su aplicación en diferentes áreas ha brindado la posibilidad de analizar grandes volúmenes de datos espaciales. Aunque están creados para manipular datos espaciales, se demanda el uso de técnicas que permitan extraer conocimiento de estos datos acumulados en Bases de Datos Espaciales (SDBMS por su nombre en idioma inglés) y el descubrimiento de patrones que sean más fáciles de entender. Se han reportado trabajos que utilizan la extracción de conocimiento automatizada mediante la minería de datos, que permite encontrar conocimiento implícito [37, 11, 38] en grandes volúmenes de datos. Sin embargo, producto de la complejidad de los tipos de datos (puntos, líneas, polígonos), los objetos y las estructuras de datos que se manejan en SDBMS, se dificulta la utilización de aproximaciones tradicionales de la minería de datos. La minería de datos espaciales provee un grupo de técnicas y herramientas para la explotación de estos datos que permiten encontrar patrones

potencialmente útiles.

En Cuba, los métodos aplicados para la estratificación en su mayoría van orientados al análisis estadístico, sin tener en cuenta la naturaleza espacial de los datos, ni el principio de la primera ley de la geografía. Se evidencia una estructuración común que inicia con análisis estadísticos apoyándose en herramientas por ejemplo: Excel, SPSS y luego se presentan los resultados en mapas temáticos utilizando herramientas SIG, por ejemplo MapInfo; lo cual reduce la eficiencia del trabajo [39, 21, 40, 2, 41, 42, 43, 44, 45, 27, 12, 46, 46, 14, 47]. Estos elementos influyen en el análisis de la relación espacial de indicadores en diferentes áreas y en la capacidad de gestión de las entidades de salud. Las medidas de similitud empleadas consideran las características con igual importancia y están enfocadas a los datos temáticos. Este tratamiento no permite describir relaciones espaciales sobre objetos y por tanto dificulta la incorporación del espacio en el proceso, favoreciendo la aparición de estratos con territorios separados, incumpliendo con la primera ley de la geografía [48] y que propicia hipótesis o modelos inexactos e inconsistentes.

A partir de la situación problemática descrita se define el siguiente **problema científico**:

el insuficiente tratamiento a la componente espacial en la estratificación de territorios limita la obtención de grupos más compactos.

El **objeto de estudio**: la minería de datos espaciales.

Para dar solución al problema se trazó el siguiente **objetivo general**: desarrollar un método de estratificación de territorios basado en Sistemas de Información Geográfica y medidas de similitud geométrica para obtener grupos más compactos.

Enmarcado en el **campo de acción**: métodos de estratificación de territorios.

El objetivo general se desagrega en los siguientes **objetivos específicos**:

1. Construir el marco teórico referencial de la investigación relacionada con la estratificación de territorios y medidas de similitud geométrica.
2. Diseñar una medida de similitud de territorio para integrar datos temáticos y datos espaciales.
3. Diseñar un método de estratificación de territorios para obtener grupos más compactos.
4. Desarrollar una herramienta computacional basada en el método propuesto para el Sistema de Información Geográfica QGIS.
5. Validar la propuesta, a través de los métodos definidos en la investigación.

Se formuló la siguiente **hipótesis**: el desarrollo de un método de estratificación de territorios basado en SIG y medidas de similitud geométrica aplicado en la herramienta informática QGIS facilitará la obtención de grupos más compactos.

Operacionalización de las variables dependientes e independientes.

Definición conceptual de la variable independiente:

Método de estratificación de territorios: es un conjunto de etapas con sus procedimiento que permite clasificar objetos en clases homogéneas a partir de analogías o relaciones que se establecen entre sus características. Definición conceptual de la variable dependiente:
Grupos más compactos: relacionado con la primera ley de la geografía y la dependencia o autocorrelación espacial, se tiene en cuenta que en el espacio todo está relacionado, pero objetos cercanos están más relacionados que objetos distantes [35, 36]. La dependencia o autocorrelación espacial aparece como consecuencia de la existencia de una relación funcional entre lo que ocurre en un punto determinado del espacio y lo que ocurre en otro lugar [49, 50, 51].
Se utilizan los siguientes **métodos de investigación**:

- Métodos teóricos: **análisis-síntesis**, para el estudio de las fuentes bibliográficas existentes relacionadas con el tema, identificando los elementos más importantes y necesarios para dar solución al problema planteado. **Histórico-lógico** para el estudio crítico de los trabajos anteriores y utilizar estos como puntos de referencia y comparación de los resultados alcanzados. **Hipotético-deductivo**, para elaborar la hipótesis de investigación y proponer líneas de trabajo a partir de resultados parciales. **Análisis documental**, con la consulta de la literatura especializada en las temáticas afines de la investigación. **Modelación**, para la representación explícita de la solución propuesta.
- Método empírico: **cuasiexperimento** para validar la propuesta se aplica el método de la presente investigación a dos casos de estudios disponibles en la literatura manteniendo los mismos indicadores y algoritmos.

La **novedad científica** se expresa en los siguientes **aportes prácticos**: un método para estratificación de territorios, medidas de similitud geométrica y XANGEO [1] sistema informático que instancia al método propuesto.
Estructura de la tesis: El presente documento está estructurada en tres capítulos:

- **Capítulo 1**: Referentes teóricos sobre la minería de datos espaciales y la estratificación de territorios. En este capítulo se presentan un conjunto de elementos que conforman los fundamentos teóricos relacionados con el objeto de estudio de la investigación. Se abordan definiciones de estratificación de territorios, los enfoques o técnicas, entre ellas: el análisis multivariado. Se caracteriza a los SIG herramienta utilizada para la geovisualización y se abordan aspectos de la minería de datos y de la minería de datos espaciales. Una de las funciones de la minería de datos es la creación de grupos (clustering en idioma inglés), basándose en algunas reglas de similitud previamente definidas. En la presente investigación estas reglas están conformadas por las medidas de similitud geométricas definidas como criterios, las cuales son abordadas en este capítulo. Se realiza un análisis crítico de algunas soluciones existentes y se describen índices de validación internos y externos para evaluar agrupamientos.

[1] Sistema para el análisis geoespacial en estudios salubristas

- **Capítulo 2**: Método de estratificación de territorios basado en Sistemas de Información Geográfica y medidas de similitud geométrica. En este capítulo se describe y fundamenta un método para la estratificación de territorios basada en SIG y medidas de similitud geométricas. El método está conformado por cinco etapas denominadas: Selección de indicadores y territorios, Preprocesamiento de indicadores, Agrupamiento, Postprocesamiento y Visualización. En cada uno de los epígrafes de este capítulo se encuentra información relacionada con las etapas. También se muestran artefactos ingenieriles resultados de las fases de análisis y diseño de la instanciación del método.
- **Capítulo 3**: Aplicación del método propuesto en la estratificación. En este capítulo se presenta la aplicación del método de estratificación de territorios basado en Sistemas de Información Geográfica y medidas de similitud geométrica a casos de estudios, con el objetivo de comprobar que se obtienen grupos más compactos. Se aplican índices internos y externos para la validación de los agrupamientos. Se muestran los resultados de la dependencia espacial utilizando la I de Moran y de las pruebas estadísticas no paramétricas aplicadas.

1. Capítulo 1: Referentes teóricos sobre la Minería de datos espaciales y la estratificación de territorios

En este capítulo se presentan los elementos que conforman los fundamentos teóricos relacionados con el objeto de estudio de la investigación. Se abordan definiciones de estratificación de territorios, los enfoques o técnicas, entre ellas: el análisis multivariado. Se detallan los componentes y características de los SIG y se abordan aspectos de la minería de datos y de la minería de datos geoespacial. Una de las tareas en la minería de datos es la creación de grupos (clustering en idioma inglés), basándose en algunas reglas de similitud previamente definidas. En la presente investigación estas reglas se definen como medidas de similitud geométricas a partir de criterios que se abordan en este capítulo. Se realiza un análisis crítico de algunas soluciones existentes y se formalizan los índices de validación internos y externos para analizar la separación que existe entre los grupos y la compacidad que hay entre las instancias que pertenecen al mismo clúster.

1.1. Estratificación de territorios: bases conceptuales

El diccionario de la Real Academia de la Lengua española registra la palabra estratificar como disponer en estratos, otra acepción se refiere a la formación o deposición de las capas [52]. En la literatura consultada no hay un consenso entre las definiciones, se aprecia la utilización de proceso o de metodología indistintamente:

- La estratificación territorial es analizada como un proceso que permite dimensionar espacialmente los eventos a través de un proceso de agregación y desagregación de los territorios a evaluar, a partir de variables seleccionadas para dichos territorios que permitan agregaciones (por homologías de las características) o desagregaciones (por heterogeneidades de estas) [21, 2].
- La estratificación es un conjunto de analogías que dan lugar a subconjuntos de unidades agregadas, denominadas estratos [13].
- La estratificación es un procedimiento que permite clasificar objetos en clases homogéneas a partir de analogías o relaciones que se establecen entre sus características [18, 11, 53, 20].
- La estratificación territorial es una metodología que permite dimensionar espacialmente los eventos a través de un proceso de agregación por homologías de las características y desagregación (por heterogeneidades de estas) de los territorios a evaluar, a partir de variables seleccionadas [12].

- Estratificación es la división en clases discretas y homogéneas de variables o conjuntos de variables que se expresan en gradientes continuos [54].

En estudios salubristas suele denominarse estratificación epidemiológica y es parte del proceso integrado de diagnóstico - intervención - evaluación [18]. La estratificación epidemiológica utiliza varios enfoques para contribuir a la selección de sitios o zonas con problemas de salud y planificar estrategias de intervención [21].
Algunos de los enfoques de la estratificación epidemiológica son: la estratificación del riesgo y del riesgo absoluto, por indicadores ponderados, los patrones de distribución de frecuencia de los principales riesgos asociados y las técnicas de análisis multivariado fundamentalmente las basadas en análisis de grupos o agrupamiento [25]. En la presente investigación se utiliza esta última porque posibilita identificar los grupos en función de las similitudes, tributando a la simplificación y representatividad del conjunto, esto se sustenta en que la estratificación es un conjunto de etapas con sus procedimientos que permite clasificar objetos en clases homogéneas a partir de analogías o relaciones que se establecen entre sus características [18, 11, 13, 53, 20].
El enfoque a utilizar en el proceso de estratificación depende de la urgencia de los resultados para tomar decisiones, la disponibilidad de información asociada al problema en estudio, los recursos disponibles, el entrenamiento de los especialistas que acometerán el estudio y su finalidad.

1.2. Análisis multivariado basado en agrupamiento

El análisis multivariado es un conjunto de técnicas de análisis de datos en expansión, incluye técnicas y métodos estadísticos. En la literatura es considerada útil porque permite al investigador extraer abundante información de los datos disponibles, preservar las correlaciones naturales entre las múltiples influencias del comportamiento y los efectos aislados de esas influencias, sin provocar el típico aislamiento de individuos o variables [55]. Se reporta su aplicación con buenos resultados en la industria, administración y centros de investigación de ámbito universitario [56, 57].
Diversas definiciones de las técnicas de análisis de datos multivariados [57, 58] reportadas en la literatura, la consideran una herramienta que tiene como principal objetivo; resumir grandes cantidades de datos por medio de pocos parámetros (simplificación) y encontrar relaciones entre: variables de respuesta y unidades experimentales.
Se puede observar que cuando existen muchas variables es posible que parte importante de la información sea redundante, en cuyo caso es necesario eliminar el exceso y dejar solo variables que tengan representatividad dentro del conjunto. Para este fin se reportan aplicaciones de las técnicas multivariadas [59]. Las técnicas multivariadas más utilizadas en el análisis de datos son: análisis de componentes principales, análisis factorial, análisis de clasificación entre los que se encuentran: discriminante, análisis multivariado de la varianza y covarianza, análisis de variables canónicas y clúster [60, 56, 59, 57].
El análisis de grupos, clasificación, conjunto o clúster (en lo adelante análisis multivariado basado en agrupamiento)

[60, 56, 59], es una técnica analítica para desarrollar subgrupos significativos de individuos u objetos. Tiene por objeto agrupar elementos en grupos homogéneos, en función de las similitudes o diferencias entre ellos. De forma específica, el objetivo del análisis multivariado basado en agrupamiento es clasificar una muestra de entidades (personas u objetos) en un número pequeño de grupos mutuamente excluyentes basados en similitudes entre las entidades. En este, a diferencia del análisis discriminante, los grupos no están predefinidos.

Habitualmente, el análisis multivariado basado en agrupamiento implica al menos dos etapas [56]. La primera es la medida de alguna forma de similitud o asociación entre las entidades para determinar cuántos grupos existen en realidad en la muestra. La segunda etapa es describir las personas o variables para determinar su composición. Este paso puede llevarse a cabo aplicando el análisis discriminante a los grupos identificados por la técnica clúster.

El análisis multivariado basado en agrupamiento construye conglomerados, de tal forma que los objetos del mismo conglomerado son más parecidos entre sí que a los objetos de otros conglomerados. Lo que se intenta es maximizar la homogeneidad de los objetos dentro de los conglomerados mientras que a la vez se maximiza la heterogeneidad entre los agregados [56]. Es la única técnica multivariante que no estima el valor teórico empíricamente, sino que utiliza el valor teórico especificado por el investigador. Su objetivo fundamental es la obtención de un conjunto de objetos en dos o más grupos basándose en su similitud para un conjunto de características especificadas (valor teórico del análisis clúster) [56]. Atendiendo a estas características es considerado idóneo para la presente investigación.

Las técnicas de agrupamiento se registran como unas de las más utilizadas y con mejores resultados para realizar la clasificación de los datos en los procesos de estratificación de territorios [20]. Los métodos multivariados son un conjunto de técnicas que tienen un carácter exploratorio y no tanto inferencial. En la literatura consultada se observa el uso de los métodos multivariados en la minería de datos (data mining en idioma inglés) [57], en la presente investigación se utiliza el análisis multivariado basado en agrupamiento.

1.3. Sistemas de Información Geográfica (SIG) en la estratificación

El desarrollo de los SIG o GIS (por sus siglas en inglés) se ha visto favorecido por la necesidad creciente de gestionar y analizar información geográfica [61, 62, 63, 64]. Su alcance se ha redefinido, incorporando nuevas funcionalidades que dan soporte a estudios más específicos, entre los que se destacan los estudios de salud [65, 66]. Los SIG son una herramienta de gran impacto por su aplicación en ramas como la agricultura [67], la meteorología [68], el turismo [69] y la salud pública [70, 71]. Se ha utilizado en los espacios docentes como instrumento científico-metodológico [72] para la identificación geográfica de establecimientos de salud, grupos de población que presentan mayor riesgo de enfermar o de morir prematuramente y por tanto que requieran mayor atención preventiva, curativa o de promoción de salud [73, 74, 75].

Entre los antecedentes de aplicación SIG sobre datos de salud, se destaca la representación de la distribución

espacial de las enfermedades, de gran interés para mostrar geográficamente las tasas de incidencia con objetivos puramente descriptivos. La representación espacial también ha sido utilizada para formular hipótesis relacionadas con la etiología de enfermedades y documentar o establecer el marco de estudios de la epidemiología [76, 13].

Varios autores han propuesto definiciones de SIG:

- Un SIG es un sistema computacional para la entrada, manejo (almacenamiento y recuperación de información), manipulación, análisis y representación de datos geográficos [77].
- Un SIG es un sistema para capturar, almacenar, comprobar, integrar, manipular, analizar y visualizar datos que están espacialmente referenciados a la tierra [78].
- Un SIG es un sistema automatizado para la captura, almacenamiento, composición, análisis y visualización de datos espaciales [79].
- Un SIG es un sistema de hardware, software y procedimientos diseñados para soportar la captura, gestión, manipulación, análisis, modelado y visualización de datos espacialmente referenciados para resolver problemas complejos de planeamiento y gestión [80].
- (. . .) GIS es una herramienta muy poderosa (. . .), se utiliza para organizar, analizar, visualizar y compartir todo tipo de datos e información de diferentes períodos históricos y en diversas escalas de análisis [81].
- Un SIG se puede considerar como la unión de componentes físicos y lógicos que permiten la gestión de información georreferenciada [82].
- SIG es un sistema que integra tecnología informática, personas e información geográfica, y cuya principal función es capturar, analizar, almacenar, editar y representar datos georreferenciados [61].
- Los SIG constituyen ser herramientas básicas para la confección de mapas digitales y para los análisis geoespaciales, en todas las esferas del saber, que van más allá de análisis estadísticos y que tributan a una mejor planificación de infraestructura, en estudios demográficos, análisis de vías de transporte, distribución de recursos, distribución y comportamiento de enfermedades en salud [36].
- Los GIS se pueden definir como los sistemas de información que procesan datos espaciales. Los datos espaciales se presentan en diferentes capas que forman el mapa, los edificios, las entidades y sus límites [38].
- SIG es un marco para recopilar, gestionar y analizar datos. Enraizado en la ciencia de la geografía, integra muchos tipos de datos. Analiza la ubicación espacial y organiza capas de información en visualizaciones utilizando mapas y escenas en tres dimensiones. Revela conocimientos más profundos sobre los datos, como patrones, relaciones y situaciones, lo que ayuda a los usuarios a tomar decisiones más inteligentes [83].

Las definiciones SIG han evolucionado para destacar el papel de la diseminación de los datos, como función ineludible de estos en ambientes distribuidos y globales de acceso a los datos [84].

La ventaja del uso de los SIG está dada por la representación gráfica de la información almacenada en bases de datos geoespaciales, que puede ser comparada y visualizada en mapas, con el dinamismo asociado a las escalas y

los cálculos geográficos [85, 86]. Para justificar su importancia y el papel que estos juegan, es necesario mencionar que se acepta por la comunidad científica que entre el 80 y 90 por ciento de la información utilizada en la toma de decisiones tiene asociada una componente espacial [21, 87].

El desarrollo de herramientas para el análisis de la información, considerando los datos relativos a su posición en el espacio, constituyen una oportunidad para las nuevas investigaciones. Los datos espaciales se pueden clasificar en datos vectoriales y ráster. Los datos vectoriales proporcionan representación a los lugares y entidades como puntos, líneas y polígonos propiciando la utilización de funciones matemáticas. El punto es el lugar exacto del mapa para la entidad, la línea representa la relación entre estas entidades y el polígono representa relaciones más complejas. Por otro lado, los datos ráster dividen el mapa en celdas adyacentes, lo que proporciona una capacidad de almacenamiento para almacenar todos los detalles en el mapa [38].

El uso de los SIG cada día tiene mayor utilidad en la salud, su empleo contribuye al fortalecimiento de la capacidad de análisis en epidemiología. Utilizándose en el análisis de la situación de salud, la vigilancia de eventos, el estudio epidemiológico, la planeación-evaluación de estrategias por zonas de salud y en la toma de decisiones [88]. Facilitan la ubicación geográfica de establecimientos de salud y grupos de población para monitorear la dirección de enfermedades.

Para el análisis de salud, se hace necesario conocer las características de cada unidad territorial, así como sus grupos poblacionales [21], a partir de diferentes indicadores, que pueden ser: demográficos, socio-económicos y ambientales. Todos estos elementos tienen un impacto determinante en la caracterización de un territorio y constituyen la base en el establecimiento de la estratificación territorial, lo cual ya fue mencionado en el epígrafe 1.1.

La utilización de los SIG en el análisis de la distribución espacial de enfermedades ha aumentado considerablemente, sustentado en las herramientas de análisis existentes que posibilitan resolver problemas asociados a la distribución espacial [35]. Sin embargo, estas no son extensibles, su utilización se limita a llevar información a la cartografía, y la componente espacial de los datos no es explotada en su totalidad. Si bien el espacio es un elemento importante en estos estudios, no siempre se le da la importancia requerida, motivado por: acceso limitado a los SIG por los costos que ellos implican, poco conocimiento de las herramientas y el tiempo de formación en el área de los SIG es elevado [11].

En la literatura consultada se encuentran la utilización de softwares SIG, como por ejemplo: gvSIG[1], ArcView[2], MapInfo[3] y QGIS[4] [2, 42, 43, 89, 45, 36, 13, 47, 90]. En esta investigación se utiliza QGIS, destacándose por su licencia GNU. Es una aplicación escritorio, con un entorno sencillo, amigable, muy intuitivo y fácil de utilizar [90].

[1] http://www.gvsig.com/es

[2] https://arcviewgroup.com/

[3] https://www.pitneybowes.com/us/location-intelligence/geographic-information-systems/mapinfo-pro.html

[4] https://www.qgis.org/en/site/

Además es multiplataforma, posibilita conexión a base de datos PostgreSQL y PostGIS. Permite la incorporación de nuevos módulos y funcionalidades implementadas en C++ y Python [91], manipula formatos ráster y vectoriales a través de las bibliotecas GDAL y OGR, así como bases de datos. Cuenta con una comunidad activa para su soporte [3].

Hay evidencias de utilización de los SIG en Cuba en algunos sectores, por ejemplo en la salud pública con el análisis y distribución de los problemas de salud, modelos epidemiológicos, fundamentalmente con enfoque a la estadística médica [3]. La acumulación de información espacial producto del desarrollo de los sistemas informáticos y en especial de los SIG, propicia la aplicación de técnicas de minería de datos espaciales para soporte a la toma de decisiones y para estudios de diagnóstico-intervención-evaluación en salud.

1.4. La minería de datos espaciales

El desarrollo alcanzado en la informática y en los sistemas informáticos, ha propiciado la utilización de la componente espacial en sus investigaciones y en el diseño de estrategias personalizadas de prevención y control en el área de la salud y la epidemiología [11]. La amplia difusión de información espacial producto del desarrollo de los SIG ha favorecido la explotación de los datos con el objetivo de encontrar conocimiento de manera automatizada. La complejidad de los tipos de datos existentes en SDBMS y las estructuras de datos que las soportan limitan la utilización de técnicas tradicionales de minería de datos, lo que propicia la aparición de nuevas técnicas que de conjunto forman la minería de datos espaciales.

La minería de datos es considerada un proceso de recopilación, búsqueda y análisis de información relevante en grandes volúmenes de datos, para descubrir patrones o relaciones [92, 93, 94, 95]. La minería de datos espaciales es considerada una rama de la minería de datos, definida como el proceso automático o semiautomático de seleccionar, explorar, modificar, visualizar y valorar grandes volúmenes de datos espaciales con el objetivo de descubrir conocimientos [11].

El análisis espacial constituye una metodología útil para la gestión de riesgos, vigilancia epidemiológica del entorno social [96, 97] y soporte a la toma de decisiones. Es un reflejo de los desafíos presentados para integrar diversas fuentes de datos en la era de BigData [98, 99]. Tiene como objetivo automatizar el proceso de descubrimiento del conocimiento [100], revelar patrones y anomalías sobre el comportamiento espacial de variables o eventos que serán validados por expertos del área estudiada [101, 102, 103].

El descubrimiento del conocimiento o patrones en bases de datos espaciales a través de la minería de datos es complejo, pues no solo se encarga de datos no espaciales, sino que tiene en cuenta la localización de los objetos y sus relaciones topológicas [35].

En la literatura consultada [104, 105, 106, 103, 107, 108, 109, 110, 64] se clasifican los métodos de minería de datos espaciales en cinco grupos:

1. Métodos basados en generalización: requieren la implementación de jerarquías de conceptos, en el caso de las base de datos espaciales estas jerarquías pueden ser temáticas o espaciales. Se recolectan sus características más importantes para la búsqueda, se caracterizan por regiones y se agrupan como datos no espaciales generalizados. Para el caso de los espaciales esta generalización puede ser presentada como la partición en regiones y su posterior fusión dependiendo de los atributos espaciales de los datos.
2. Métodos de reconocimiento de patrones: pueden ser usados para realizar reconocimientos y categorizaciones automáticas de fotografías, imágenes y textos.
3. Métodos explorando asociaciones espaciales: permiten descubrir reglas de asociación espacial, es decir reglas que asocien a uno o más objetos espaciales, con otro u otros objetos espaciales XY (donde XyY son un conjunto de predicados espaciales o no) [111].
4. Métodos usando aproximación y agregación: utilizan aproximación y agregación para descubrir conocimiento sobre la base de las características representativas del conjunto de datos [103].
5. Métodos usando agrupamiento: son utilizados para crear agrupaciones o asociaciones de datos, teniendo en cuenta cuestiones de similaridad. Los algoritmos de agrupamiento han sido empleados en reconocimiento del habla, en segmentación de imágenes, visión por computador [112], en minería de datos para extraer conocimiento desde fuentes de datos, en la recuperación de información, en minería de textos [113, 114, 115, 116] y en minería de datos geoespaciales [117, 118, 11, 111].

Desde el enfoque de la minería de datos geoespaciales destaca la utilización de los algoritmos de agrupamiento en estudios estratificados. Una clasificación general los divide en tres grupos fundamentales: particionales, jerárquicos y basados en densidad [104, 105, 119, 120, 106, 103, 107, 108, 109, 103, 110, 64] o categorías según el procedimiento que utilizan para agrupar los objetos [121, 122], como se muestra en la figura 1.1:

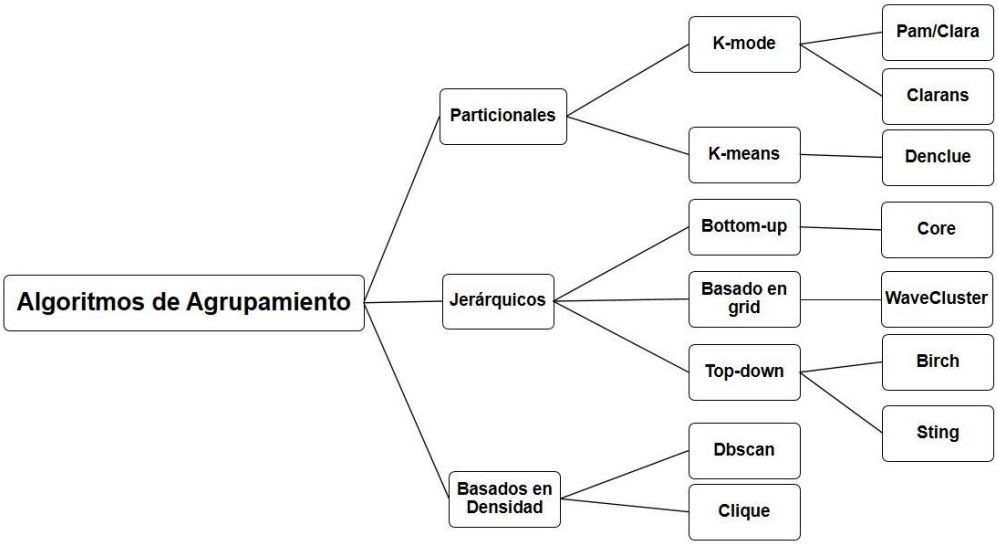


Figura 1.1: *Algoritmos de agrupamiento, tomado de[1].*

La figura 1.1 muestra la relación entre los diferentes tipos de algoritmos de agrupamiento espacial, aunque se debe tener en cuenta que existen otras clasificaciones reportadas en la literatura [119, 123, 95, 124].

1. Agrupamiento por particiones (Partitional clustering en idioma inglés): incluye a los algoritmos que crean particiones de los datos, de tal forma que los objetos en una partición (clúster) sean más similares entre sí que objetos de otras particiones [125, 126, 127]. Los algoritmos de agrupamiento de particiones dividen las instancias de los datos en k particiones, donde cada partición representa un grupo. La partición es realizada en función de cierta función objetivo, los grupos se forman para optimizar el criterio de partición objetiva. Esta función puede ser de disimilitud basada en la distancia, para que los objetos dentro de un clúster sean "similares", mientras que los objetos de diferentes grupos son diferentes [128, 129, 124]. En esta clasificación se encuentran los algoritmos K-Means, PAM (Partition around mediods) y CLARA (Clustering Large Applications).

    - El algoritmo K-Means (por sus siglas en idioma inglés), ampliamente reportado en la literatura consultada es considerado un algoritmo simple, eficiente [130, 129, 131] y con buenos resultados en estudios salubristas [132, 2, 133, 134, 131]. Selecciona a los representantes iniciales de forma aleatoria o a partir de una heurística. Luego agrupa los objetos según su proximidad entre sí según la función de distancia definida. Se reporta el uso de la distancia euclidiana que reduce el ruido y los valores atípicos. Sigue una forma sencilla para dividir una base de datos dado el valor k, que define la cantidad de grupos (fijados a priori), donde cada grupo tiene asociado un centroide (centro geométrico del grupo). Los datos se asignan al grupo cuyo centroide esté más cerca, utilizando la distancia definida y luego itera para actualizar los centroides en función de las asignaciones. Una ventaja de este algoritmo es su simplicidad. Como desventaja es muy difícil especificar el número de conglomerados por adelantado, también es sensible a valores atípicos. Otro inconveniente es que los centríodos no son significativos en la mayoría de los problemas [128, 129, 130]. Sin embargo es reportado en la literatura consultada como unos de los más utilizados y con mejores resultados en estratificaciones en salud [132, 2, 133, 134, 131].
    - El algoritmo PAM (por sus siglas en idioma inglés) fue presentado por Kaufman y Rousseeuw. Está basado en k objetos representativos, llamados medoides, entre los objetos del conjunto de datos. Los medoides son puntos con el promedio más pequeño de semejanza a todos los otros puntos. El algoritmo sigue los mismos pasos que son seguidos por el algoritmo K-Means, pero el uso de medoides en lugar de medios hace que el algoritmo sea más robusto a los valores atípicos. PAM también se puede usar en conjuntos de datos que tienen datos discretos u otros tipos, como datos binarios. Uno de los problemas del algoritmo PAM es que la cantidad deseada de grupos debe ser predeterminado [128, 129].
    - El algoritmo CLARA (por sus siglas en idioma inglés) es un método basado en PAM que intenta tratar con aplicaciones a grandes conjuntos de datos. CLARA usa el algoritmo PAM para agrupar una

muestra de un conjunto de objetos en k subconjuntos. Después de este primer paso, cada objeto que no pertenece a la muestra inicial se asigna al objeto representativo más cercano y se obtiene una medida de agrupamiento de todo el conjunto de datos. Esta medida se compara con otras medidas obtenidas a partir de la aplicación del algoritmo en n muestras iniciales diferentes. La mejor agrupación obtenida de las diferentes muestras es la seleccionada por el algoritmo [128, 129].

- El algoritmo CLARANS (por sus siglas en idioma inglés) combina las técnicas de muestreo con PAM. El proceso de agrupamiento se puede describir a partir de como buscar un grafo donde cada nodo sea una solución potencial, es decir, un conjunto de k medoides. El agrupamiento obtenido después de reemplazar a un medoide se denomina vecino de la agrupación actual. CLARANS selecciona un nodo y lo compara con un número definido por el usuario de sus vecinos buscando un mínimo local. Si se encuentra un mejor vecino, que tiene un error de cuadrado menor, CLARANS se mueve al nodo del vecino y el proceso inicia de nuevo; de lo contrario, la agrupación actual es un óptimo local. Si se encuentra el óptimo local, CLARANS comienza con un nuevo nodo seleccionado al azar en busca de un nuevo óptimo local [128, 129].

2. Agrupamiento jerárquico en idioma inglés (Hierarchical clustering), como su nombre indica, construyen una jerarquía de grupos, uniendo o dividiendo los grupos de acuerdo a una cierta función de similaridad/disimilaridad entre los grupos [127, 135, 136]. Los algoritmos de agrupamiento jerárquico dividen o fusionan un conjunto de datos en una secuencia de particiones anidadas. La jerarquía de las particiones anidadas puede ser aglomerativa (ascendente) o divisiva (descendente). En el método aglomerativo, la agrupación comienza con cada objeto individual en un único grupo y continúa agrupando los pares más cercanos de grupos hasta que todos los objetos estén juntos en solo un grupo. La agrupación jerárquica divisiva, por otro lado, comienza con todos los objetos en un solo grupo y sigue dividiendo los grupos más grandes en grupos más pequeños hasta que todos los objetos se separan en grupos de unidades [129]. Los métodos jerárquicos generan un dendrograma como una estructura de árbol que representa el proceso de agrupamiento [124]. Algunas desventajas de los métodos jerárquicos es que no se puede retroceder o deshacer ningún paso, no hay una función objetivo que se minimice como en los métodos de partición, no existe un criterio de parada y luego de construidos los clúster, no vuelven a ser visitados para mejorarlos. Estos métodos requieren $O(n^2)$.
3. Basado en cuadrículas del idioma inglés (Grid-based), los objetos de datos se dividen en cuadrículas. Estos en su mayoría son más simples que la interpolación; los enfoques estocásticos abarcan los métodos geoestadísticos. Se basan en una estructura de rejilla de múltiples niveles. Todo el espacio se cuantifica en un número finito de celdas en las que se realizan operaciones para la agrupación. La información resumida sobre el área cubierta por cada celda se almacena como un atributo de la celda. Su principal ventaja es su tiempo de procesamiento. Sin embargo, la información resumida conduce a la pérdida de información. La

agrupación basada en rejillas reduce significativa la complejidad computacional, especialmente para agrupar conjuntos de datos muy grandes. El enfoque de agrupamiento basado en rejillas difiere de los algoritmos de agrupamiento convencionales en que no se refiere a los puntos de datos, sino al espacio de valores que rodea los puntos de datos. Wave-Cluster y STING son ejemplos de algoritmos de esta categoría [120, 123].

4. Basado en la densidad en idioma inglés (Density-based), aquí los objetos de datos están separados en subregiones de densidad, conectividad y límite, estrechamente relacionados con los puntos más cercanos. Un clúster definido como un componente denso conectado, crece en cualquier dirección a la que conduce la densidad. Los algoritmos basados en densidad son capaces de descubrir grupos de formas arbitrarias, esto proporciona una protección natural contra valores atípicos. Por lo tanto, se analiza la densidad general de un punto para determinar las funciones de los conjuntos de datos que influyen en un punto, un algoritmo de esta clasificación es DBSCAN [120, 123].

    - DBSCAN (por sus siglas en idioma inglés) conecta objetos núcleos y sus vecinos para formar regiones densas como clúster [137, 138]. Puede descubrir grupos de forma arbitraria y también maneja valores atípicos de manera efectiva, obtiene grupos al encontrar el número de puntos dentro de la distancia especificada desde un punto dado [137]. Ampliamente utilizado en muchas áreas de la ciencia debido a su simplicidad y la capacidad de detectar grupos de diferentes tamaños y formas [139]. Los resultados del agrupamiento final depende del orden en que se procesan los objetos en el transcurso del algoritmo ejecutado.

En la minería de datos geoespaciales destaca la utilización de los algoritmos de agrupamiento, la elección de dicho algoritmo depende de varios factores, por ejemplo: tipo de datos disponibles, criterio de agrupamiento, complejidad, detección de valores atípicos y el propósito particular [128, 129]. En la presente investigación se utilizan los algoritmos de agrupamiento espacial porque forman grupos basándose en alguna función de distancia o similitud previamente definida en correspondencia con los objetivos que se persiguen [140, 95].

1.5. Funciones de distancia y de similitud

En la Ciencia de la Información Geográfica o Geociencia (GIScience), la distancia juega un papel importante para resolver problemas asociados al espacio y constituye la base para la recuperación e integración semántica [141, 142]. Una función $d : X \times X \to \mathbb{R}$ es una distancia o disimilitud en X si $\forall x, y \in X$ para la que se cumple que es no negativa $d(x, y) \geq 0$, simétrica $d(x, y) = d(y, x)$ y reflexiva $d(x, x) = 0$.

La similitud es una función $s : X \times X \to \mathbb{R}$ en X para la cual se cumple que es no negativa, simétrica y la desigualdad $s(x, y) \leq s(x, x) \forall x, y \in X$, solamente se cumple la igualdad cuando $x = y$.

Cuando la distancia se representa en el intervalo $[0, 1]$ se pueden utilizar las transformaciones $d = 1 - s$, $d = \frac{1-s}{s}$, $d = \sqrt{1 - s}$, $d = \sqrt{2(1 - s^2)}$, $d = \arccos s$ y $d = -\ln s$ por solo mencionar algunas.

Las definiciones de distancia o similitud a utilizar están en correspondencia con la naturaleza de los datos y los objetivos que persigue el estudio [143, 144]. Teniendo en cuenta que un caso viene descrito por un conjunto de atributos, la aproximación básica para el cálculo de la similitud consiste en contabilizar los valores iguales en los atributos comunes de los casos a comparar [145].

En tareas de agrupamiento, la similitud o disimilitud entre objetos es una medida de correspondencia, parecido o diferencia entre los objetos [56]. La similitud se expresa como una función lineal de las características según igualdad y diferencia [146]. Las características iguales tienden a incrementar más la similitud de lo que las características diferentes tienden a disminuirlas y se calcula mediante el vector distancia entre dos puntos correspondientes en ese espacio [146, 147, 148, 149, 150].

Para resolver problemas de agrupamiento ha sido ampliamente reportada la utilización de la distancia euclidiana, de Manhattan, Chebychev, Minkowski y la distancia de Mahalanobis [151]. Estas se definen de la siguiente manera: Sea x una instancia arbitrariamente descrita por el vector de características:

$$(a_1(x), a_2(x), ..., a_n(x)) \tag{1.1}$$

donde $a_r(x)$ denota el valor de r-ésimo atributo de la instancia x; entonces la distancia entre dos instancias x_i y x_j, es definida por $d(x_i, x_j)$, para cada una de las siguientes distancias:

Distancia de Manhattan:

$$d(x_i, x_j) = \sum_{i=1}^{n} |(a_r(x_i) - a_r(x_j))| \tag{1.2}$$

Distancia de Chebychev:

$$d(x_i, x_j) = max_{i=1,...n} |(a_r(x_i) - a_r(x_j))| \tag{1.3}$$

Distancia de Minkowski:

$$d(x_i, x_j) = (\sum_{i=1}^{n} |(a_r(x_i) - a_r(x_j))|^p)^{\frac{1}{p}} \tag{1.4}$$

La distancia de Minkowski es una generalización de las distancias euclídea, Manhattan y Chebychev, donde un parámetro p debe ser definido. Si $p = 1$, es la distancia de Manhattan, si $p = 2$, es la distancia euclídea y finalmente si $p = \infty$, es la distancia de Chebychev. Adicionalmente, la distancia euclídea es un caso particular de la distancia de Mahalanobis:

$$d(x_i, x_j) = ((a_r(x_i) - a_r(x_j))^t S^{-1} (a_r(x_i) - a_r(x_j)))^{\frac{1}{2}} \tag{1.5}$$

donde S es la matriz de covarianza y t elemento de la matriz.

En la distancia euclídea no se tiene en cuenta la correlación entre los atributos. La función distancia euclidiana ponderada se identifica como una de las más utilizadas y sencillas [152], con valores numéricos proporciona resultados satisfactorios en la clasificación [153]:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} w_k (x_{ik} - x_{jk})^2} \tag{1.6}$$

Donde x_i es el territorio i, x_j es el territorio j, w_k es el aporte informacional del indicador k, p total de indicadores, x_ik valor de x_i para el indicador k, para $k = 1, 2, ..., p, j = 1, 2, ...n$

Para el análisis de datos numéricos se han reportado otras medidas como la similitud de Ruzicka, Soergel y Roberts que utilizan una distancia mínima (min) entre los miembros más cercanos de los grupos, es decir, la distancia establecida de conjunto y una distancia máxima (max) entre los miembros más alejados de los grupos, es decir, la distancia de expansión [154].

Similitud de Ruzicka:

$$\frac{\sum min\{x_i, y_i\}}{\sum max\{x_i, y_i\}} \tag{1.7}$$

Esta se corresponde con la distancia de Soergel a partir de aplicar una transformación:

$$1 - \frac{\sum min\{x_i, y_i\}}{\sum max\{x_i, y_i\}} = \frac{\sum \mid x_i, y_i \mid}{\sum max\{x_i, y_i\}} \tag{1.8}$$

La similitud de Roberts se define como:

$$\frac{\sum (x_i + y_i) \frac{min\{x_i, y_i\}}{max\{x_i, y_i\}}}{\sum (x_i + y_i)} \tag{1.9}$$

Se identifican también funciones basadas en la correlación donde destacan la similitud de correlación de Pearson y similitud coseno 1.10 [124].

$$\frac{\langle x, y \rangle}{\parallel x \parallel_2 . \parallel y \parallel_2} = \cos \phi \tag{1.10}$$

Donde ϕ es el ángulo entre los vectores x y y.

Para el análisis espacial desde este enfoque se han introducido medidas asociadas a la dependencia espacial y a la autocorrelación. La dependencia espacial es una medida del grado de dependencia asociativa de mediciones de una variable en el espacio.

En los SIG la definición de similitud es importante debido a la dificultad para obtener representaciones satisfactorias de los fenómenos geográficos y a la variedad de formalizaciones que existen de sus propiedades espaciales tales como su forma, localización y relaciones espaciales [155, 150]. El modelo de datos (vectoriales y ráster) [156], define la manera en que se van a representar las características espaciales en un SIG y por tanto existe también dependencia entre el modelo de representación y la definición de similitud o distancia. Las medidas de similitud o distancia a utilizar dependen del análisis a realizar, por lo cual se pudiera tener en cuenta diferentes criterios para realizar dicha investigación.

Para los modelos vectoriales, la componente espacial es representada mediante una geometría por lo que las medidas de similitud pueden ser geométricas. Las medidas de similitud geométricas se pueden definir como un conjunto de parámetros, atributos geográficos o criterios, ligados a objetos en un mapa, que posibilita su comparación. De esta forma se puede establecer cierta similitud, en dependencia de la relación arrojada por los atributos. Dentro de los criterios se han reportado la distancia, el área, el perímetro y la densidad [157, 158].

Para la estratificación de territorios la selección de los criterios es de suma importancia, dado a que estos pueden variar en la misma cantidad que varían las operaciones matemáticas sobre los polígonos. En esta investigación se aborda la dinámica espacial desde estudios salubristas, por ello se abordan los criterios de distancia entre los polígonos que representan a los territorios, la conectividad y el criterio del tamaño [157, 158].

1.6. Análisis de las soluciones existentes

La determinación de agrupaciones espaciales de las enfermedades tiene un renovado interés sobre todo en las zonas con menos recursos. La aplicación de diversas técnicas informáticas en diferentes estudios de salud sirve como soporte a la toma de decisiones. En la bibliografía consultada se encuentran investigaciones relacionadas con la minería de datos espaciales con resultados interesantes por ejemplo: estratificaciones de territorios con relaciones entre indicadores socio-económicos y ambientales, en países como los Estados Unidos, España [159] y México [160]. En el primero se identificaron áreas de salud con bajos índices de educación, alto desempleo y trabajos mal pagados, tenían las mayores tasas de mortalidad por cáncer (colorrectal, pulmón, mama y próstata) [161].

Otros están enfocado en las distribuciones de recursos por zonas de riesgo por ejemplo: detección de esquistosomiasis en zonas pantanosas y lacustres [162], el control de la malaria en la provincia de Mpumalanga, Sudáfrica [163]. La propuesta presentada en [132], parte del programa atención integral de los servicios de salud de Hidalgo; la misma permite realizar estructuraciones de los municipios del estado de Hidalgo de forma automatizada. Brinda al usuario la opción de obtener estratificaciones libres o restringidas a un cierto número de grupos. Se basa en técnicas del Reconocimiento Lógico Combinatorio de Patrones (RLCP), permitiendo manejar diversas formas en la presentación de resultados, en forma gráfica y tabular, así como brindar información acerca de qué indicadores influyen más en la obtención del promedio de riesgo total de cada uno de los grupos formados. Como limitantes se deben especificar: la herramienta es privativa, solo permite estratificar los 84 municipios que conforman el estado Hidalgo y analiza solamente indicadores estadísticos.

El Estratificador INEGI[5] (Instituto Nacional de Estadística y Geografía) permite construir agrupaciones o estratificaciones de áreas geográficas sobre la base de información estadística. El sistema brinda al usuario la opción de seleccionar las variables que muestran mayor afinidad con el tema de su interés y elegir uno o más procedimientos de estratificación; de este modo será posible disponer de dos o más estratificaciones alternativas. Incluye, una serie de ayudas gráficas que permiten al analista realizar comparaciones y decidir cuál de todas las combinaciones de datos y métodos satisface de la mejor manera sus objetivos. Representa los resultados mediante diferentes gráficos que pueden ser: mapas temáticos, burbujas, y centroides. Como limitantes se deben especificar: utiliza para la clasificación de los territorios tres algoritmos de agrupamiento: K-medias, Mulvar y MClust, analiza solo indicadores estadísticos, no permite almacenar las estratificaciones que se realizan y brinda informaciones

[5] https://www.inegi.org.mx/temas/recursospriv/

referente a servicios, morbilidad y mortalidad hospitalaria proporcionada por los establecimientos particulares de salud en México [164].

En Cuba las estratificaciones de territorios para analizar el comportamiento de variables en el espacio se presenta como la de mayor utilización en los estudios de salud, con aportes en el descubrimiento de patrones relacionados con desigualdades socioeconómicas, eventos epidemiológicos, focos de contaminación ambiental y factores de riesgos [39, 21, 2, 41, 165, 43, 166, 45, 27, 46, 167, 36, 14, 13, 3, 168]. Permitiendo identificar las regiones donde hay que prestar especial atención en la vigilancia epidemiológica y adecuar más los planes de prevención. En su mayoría los estudios cuentan de dos partes: en una primera se realizan los análisis estadísticos con herramientas como Excel, SPSS y luego se presentan los resultados en mapas temáticos utilizando herramientas SIG, por ejemplo MapInfo [2, 43, 36, 13]

El aumento del uso de los GIS en las áreas de salud facilita la incorporación de estas técnicas a los análisis diarios por parte del personal de la salud, sin embargo se observa que se limita a llevar información a la cartografía y la componente espacial de los datos no es explotada en su totalidad.

1.7. Validación del agrupamiento utilizando índices internos y externos

Uno de los desafíos fundamentales de la agrupación es evaluar resultados, sin información auxiliar. Un enfoque común para la evaluación de los resultados de agrupamiento es utilizar la validez de índices.

Los índices internos miden la calidad de la solución en función de la distribución de las instancias por los agrupamientos, es decir, evalúan la separación que existe entre los clústeres y la compacidad que hay entre las instancias que pertenecen al mismo clúster. Este tipo de índice es el único que se puede aplicar cuando el conjunto de datos no aporta ningún dato adicional. Por otra parte, los índices externos son aquellos que evalúan los agrupamientos en función de algún atributo externo como puede ser la clase. Los índices de este tipo comparan el resultado del agrupamiento con el de una solución global denominada ground truth. De esta forma los índices saben a priori la solución óptima, así como el número óptimo de clústeres del conjunto de datos, ya que el ground truth contiene esta información.

En la literatura se encuentran diferentes tipos de índices, la selección depende del tipo de información [169, 170, 171, 172]. Los índices de validez de agrupamiento suelen definirse combinando compacidad (mide la cercanía de los elementos del clúster, una medida común de la compacidad es la varianza) y separabilidad (calcula la distancia entre dos grupos diferentes, indica que tan distintos son los dos grupos).

La literatura reporta tres enfoques para estudiar la validez del agrupamiento. El primero se basa en criterios externos. Esto implica evaluar los resultados de un algoritmo de agrupamiento con información que no está contenida en el conjunto de datos. El segundo enfoque se basa en criterios internos. Evaluar los resultados de un algoritmo de agrupamiento usando información que involucre los vectores del conjuntos de datos. Los criterios

se pueden dividir en dos grupos: el que evalúa el ajuste entre los datos y la estructura esperada y otros que se centran en la estabilidad de la solución. El tercer enfoque se basa en criterios relativos, consiste en evaluar los resultados (estructura de agrupación) comparándolos con otros esquemas de agrupación.

Los índices externos pueden clasificarse en: set matching (establecer coincidencia), pair-counting (conteo de pares) y information theory (teoría de la información) [172]. La primera clasificación establece que la etiqueta de cada instancia se corresponde con un clúster por ejemplo: Purity (Pureza), F-measure y Goodman-Kruskal. Los índices pair-counting se basan en la comparación entre el número de instancias con la misma etiqueta y el resultado del clúster, en esta categoría se encuentran: rand index, adjusted rand index, Jaccard, Fowlkes-Mallows, Hubert Statistic y Minkowski score [173]. Entre los índices basados en información está la entropía (Entropy en idioma inglés).

- Entropy mide la pureza de las etiquetas de clase de los grupos. Para el caso en que todos los grupos están formados por objetos con una sola clase etiqueta, la entropía es 0; a medida que las etiquetas de clase de los objetos en un grupo se vuelven más variadas, la entropía aumenta. Para calcular la entropía se debe calcular la distribución de clase de los objetos en cada grupo:

$$E_j = \sum_i p_{ij} log(p_{ij}) \tag{1.11}$$

  La Entropía total para un conjunto de grupos se calcula como la suma ponderada de las entropías de todos los grupos:

$$E = \sum_{j=1}^{m} \frac{n_j}{n} E_j \tag{1.12}$$

  Donde: $n_j$es el tamaño del clúster j,m es el número de clústeres, n es el número total de puntos.
- Pureza (en idioma inglés Purity) calcula la pureza de un conjunto de agrupaciones:

$$P_j = \frac{1}{n_j} Max_j(n_j^i) \tag{1.13}$$

  Donde: j es el número de objetos en j con la etiqueta de clase i. P_j es una fracción del tamaño total del grupo que la clase más grande de los objetos asignados a ese grupo representan. La pureza total de la solución de agrupamiento se obtiene como una suma ponderada de la purezas de los grupos individuales y se calcula:

$$Pureza = \sum_{j=1}^{m} \frac{n_j}{n} P_j \tag{1.14}$$

  Donde: $n_j$es el tamaño del clúster j,m es el número de clústeres, n es el número total de objetos. La Pureza mide el hecho de que un clúster contenga solo una clase entre sus datos, se obtienen valores en intervalo de [0-1].
- F-measure combina los conceptos de precisión y recuperación de la información. La recuperación y la precisión de un grupo para cada clase se calcula:

$$Recall(i, j) = \frac{n_{ij}}{n_i} \tag{1.15}$$

y

$$Precision(i,j) = \frac{n_{ij}}{n_i} \tag{1.16}$$

Donde: n_{ij} es el número de objetos de la clase i que están en el grupo j, n_j es el número de objetos en el grupo j, y n_i, es el número de objetos en la clase i. La F-measure del clúster j y la clase i viene dada por la siguiente ecuación:

$$F(i,j) = \frac{2n_{ij}}{n_i} \tag{1.17}$$

Los valores están dentro del intervalo [0-1]. Los valores más grandes indican una mayor calidad de agrupamiento.

- El índice de Folkes-Mallows se define como:

$$C = \frac{yy}{\sqrt{(yy+yn)*(yy+ny)}} \tag{1.18}$$

Este índice es la media geométrica de los coeficientes de precisión y recuperación:

$$C = \sqrt{P}R \tag{1.19}$$

Un alto valor de este índice significa una mejor precisión.

- Índice Jaccard se define como:

$$C = \frac{yy}{(yy+yn+ny)} \tag{1.20}$$

Los valores del índice Jaccard oscilan entre [0-1]. Los valores mayores indican la mejor validez del clúster. Los valores se interpretan: cero si no hay elementos que intercepten y uno si todos los elementos interceptan.

- Rand index se define como:

$$C = \frac{yy+nn}{N_t} \tag{1.21}$$

Los valores del índice varían entre [0-1]. Mayor valor indica que todas las instancias de datos son agrupado correctamente y el clúster contiene sólo instancias puras.

- Precisión se define como:

$$P = \frac{yy}{yy+ny} \tag{1.22}$$

Se denota por yy, yn, ny, nn (y significa sí, y n significa no) el número de puntos que pertenecen a estas cuatro categorías, respectivamente. N_t es el número total de pares de puntos [174].

El procedimiento general para determinar la mejor partición y el número de grupo óptimo de un conjunto de objetos mediante el uso de medidas de validación interna es el siguiente: Paso 1: Inicialice una lista de algoritmos de agrupamiento que se aplicarán al conjunto de datos. Paso 2: para cada algoritmo de agrupamiento, use diferentes combinaciones de parámetros para obtener diferentes resultados de agrupamiento. Paso 3: Calcule el índice de

validación interna correspondiente de cada partición obtenida en el paso 2. Paso 4: Elija la mejor partición y el número de clúster óptimo de acuerdo con los criterios.

La literatura consultada recomienda en los índices de validación internos a: Silhouette, Davies-Bouldin y Calinski-Harabasz [173], los cuales son utilizados en la presente investigación.

- Silhouette para un grupo dado, $X_j(j = 1, ...c)$, la técnica de silueta asigna a la i muestra de la medida de calidad X_j, $s(i) = (i = 1, ...m)$, conocida como el ancho de la silueta. Este valor es un indicador de confianza en la membresía de la muestra i en el clúster X_j y se define como:

$$s(i) = \frac{(b(i) - a(i))}{Max\left\{a(i), b)(i)\right\}} \tag{1.23}$$

  Donde: $a_(i)$ es la distancia promedio entre la muestra i y todas las muestras incluidas en X_j; $b_(i)$ es la distancia promedio mínima. El valor de $s(i)$ puede variar entre:-1 y 1. –1=mal agrupamiento, 0=indiferente, 1=bueno.

  Este índice valida el rendimiento del agrupamiento en función de la diferencia entre pares de distancias y dentro del agrupamiento. El número de clúster óptimo se determina al maximizar el valor de este índice.

- Davies Bouldin (DB) este índice identifica conjuntos de clústeres que son compactos y bien separados. Emplea como medida de compacidad de un clúster la media de las distancias de sus puntos a su centroide, mientras que como medida de separabilidad utiliza la distancia entre los clústeres, que puede ser, por ejemplo, la distancia euclídea entre los centros.

$$BD = \frac{1}{c}\sum_{i=1}^{c} Max_{i\neq j}\left\{\frac{d(X_i) + d(X_j)}{d(c_i, c_j)}\right\} \tag{1.24}$$

  Donde c denota el número de grupos i, j son etiquetas de grupo, $d(X_i)$ y $d(X_j)$ son todas muestras en los grupos i y j a sus respectivos centroides del grupo, $d(c_i, c_j)$ es la distancia entre estos centroides. Un valor menor de DB indica una mejor solución de agrupamiento. Para escoger el número de clústeres adecuado se toma el valor c que minimiza el índice de Davies Bouldin porque eso significa que los clústeres son más compactos y están más separados. Es decir cuanto más pequeño es el índice, mejor es el resultado del agrupamiento. Al minimizar este índice, los grupos son los más distintos entre sí, y por lo tanto, logran la mejor partición.

- Índice de Calinski-Harabasz (CH) ha sido utilizado para encontrar el número óptimo de agrupamientos, comparándolo con otros índices. Este índice se calcula:

$$\frac{trace(S_B)}{trace(S_W)} * \frac{n_p - 1}{n_p - k}) \tag{1.25}$$

  Donde: S_B es la matriz de dispersión entre grupos, S_W la matriz de dispersión interna, n_p el número de muestra agrupada y k el número de grupos. El índice de Calinski-Harabasz evalúa la validez del

conglomerado basándose en la suma de cuadrados promedio entre y dentro del conglomerado. El índice mide la separación en función de la distancia máxima entre los centros del clúster y la compacidad basada en la suma de las distancias entre los objetos y su centro. A mayor valor mejor solución.

1.8. Conclusiones del capítulo

La construcción del marco teórico referencial de la investigación, relacionado con la estratificación de territorios generó las siguientes conclusiones:

- La estratificación permite clasificar objetos en clases homogéneas a partir de analogías o relaciones y es muy utilizada en estudios de salud. En Cuba los trabajos reportados presentan limitaciones, pues se basan en análisis de indicadores estadísticos sin tener en cuenta la naturaleza espacial de los datos, lo que contradice la primera ley de la geografía.
- El uso de los Sistemas de Información Geográfica ha aumentado considerablemente, sin embargo su utilización en el sector de la salud aún se limita a la visualización de mapas y no se explotan en su totalidad la componente espacial de los datos.
- La definición del marco teórico referencial de la investigación relacionado con el proceso de estratificación de territorios, fundamentó la necesidad de implementar medidas de similitud geométricas entre polígonos para favorecer la incorporación de la componente espacial en el proceso de estratificación de territorios.
- La definición de criterios de similitud entre polígonos en función de las características de estudios de espacialidad en salud constituye una tarea importante que garantiza el diseño de medidas que integren la componente espacial y describan relaciones espaciales sobre objetos.
- Las técnicas de agrupamiento se registran como las más utilizadas y con mejores resultados para realizar la clasificación de los datos en procesos de estratificación de territorio. El estudio de los algoritmos de agrupamiento permitió identificar al algoritmo K-Means como el más utilizado en estos estudios.

2. Capítulo 2: Método de estratificación de territorio basado en Sistemas de Información Geográfica y medidas de similitud geométrica

En este capítulo se describe y fundamenta un método para la estratificación de territorios basada en SIG y medidas de similitud geométricas. Se presenta el paradigma empleado para ejecutar la investigación y los artefactos resultantes a partir de la ejecución de las actividades definidas para las ciencias del diseño. El método está conformado por cinco etapas que integran los procedimientos identificados en la literatura para estudios de este tipo. Se presenta XANGEO como solución informática que instancia al método.

2.1. Paradigma utilizado en el desarrollo del método

La presente investigación está enmarcada en la disciplina de los sistemas de información (SI). Los SI son esencialmente artefactos que capturan y representan el conocimiento sobre ciertos dominios. Los paradigmas que sustentan la investigación en la disciplina de los SI se identifican como: las ciencias del comportamiento y las ciencias del diseño [175]. El primer paradigma busca el desarrollo y verificación de teorías que expliquen o pronostiquen el comportamiento humano u organizacional. Por su parte las ciencias del diseño abordan la creación de soluciones innovadoras, capacidades tecnológicas y productos que faciliten el análisis, diseño, implementación, gestión y uso de los sistemas de información de manera efectiva y eficiente [176]. Desde este paradigma se crean y evalúan artefactos que intentan resolver problemas identificados [177].

A partir de la naturaleza del problema abordado en esta investigación y la relación que existe entre su campo de acción y la disciplina de los SI, la propuesta se desarrolla bajo el paradigma de las ciencias del diseño. Desde este enfoque se debe producir un artefacto viable en la forma de un constructo, un modelo, un método o una instanciación. Los constructos constituyen el vocabulario conceptual de un dominio a partir del cual se pueden definir y comunicar el problema en cuestión y su solución.

Los modelos representan el problema y solución a partir de un conjunto de proposiciones o sentencias que expresan relaciones entre constructos. Los métodos proveen guías sobre cómo resolver problemas y encontrar las soluciones, pueden expresarse a partir de algoritmos, descripciones textuales del proceso de búsqueda de las soluciones o combinaciones de ambas. Finalmente las instanciaciones muestran la viabilidad de implementación de constructos, modelos y métodos a partir de su operacionalización, lo que facilita la evaluación concreta del artefacto que se

instancia.

Los constructos disponibles dentro del objeto de estudio que se aborda en la presente investigación son suficientes para describir adecuadamente el problema que se aborda. Además los enfoques aportados en investigaciones precedentes con relación a estudios estratificados son válidos aunque diversos y con limitaciones para integrar la componente espacial de los datos en el espacio de solución del problema. Por lo que se hace necesario introducir una propuesta que integre los enfoques disponibles para explotar el espacio de solución teniendo en cuenta la componente espacial. Se propone un método para la estratificación de territorios que combina algoritmos y descripciones textuales e integra criterios de similitud geométricos con el objetivo de obtener estratos más compactos. Se implementa una instanciación del método propuesto para evaluar su viabilidad.

La investigación se planificó y ejecutó a partir del proceso definido en [178] para investigaciones bajo el paradigma de las ciencias del diseño y que define las etapas: identificación del problema y motivación, objetivos de la solución, diseño y desarrollo, demostración, evaluación y comunicación. Los elementos de las dos primeras etapas fueron presentados en la introducción de este documento. Las etapas de diseño y desarrollo abarcan la creación de los artefactos asociados a la solución e incluye la sistematización del vocabulario conceptual disponible y la identificación de los requisitos.

El vocabulario conceptual y los requisitos fueron determinados a partir de los referentes analizados en el capítulo 1. Los artefactos creados serán descritos en los siguientes epígrafes. La demostración y evaluación de su desempeño en la solución del problema serán analizados en el capítulo 3. Como evidencias de la comunicación a la comunidad científica, los principales aportes de este trabajo se encuentran publicados en artículos de revistas y presentados en diferentes conferencias científicas.

2.2. Método de Estratificación de territorios basado en Sistemas de Información Geográfica y medidas de similitud geométrica. Aspectos generales

El aporte fundamental de esta investigación es un método de estratificación de territorios basado en SIG y medidas de similitud geométrica que da continuidad al desarrollo de herramientas y técnicas para el análisis espacial en estudios salubristas. El método está conformado por cinco etapas que cubren los procedimientos identificados en la literatura para este tipo de estudio. Las etapas propuestas se basan en el enfoque de análisis de datos geoespaciales y se denominan: Selección de indicadores y territorios, Preprocesamiento de indicadores, Agrupamiento, Postprocesamiento y Visualización. El objetivo de este método es: caracterizar entidades administrativas a partir de la estratificación basada en indicadores y la evaluación del riegos como soporte metodológico a la toma de decisiones en estudios salubristas.

Se sustenta en los siguientes principios:

- **Integración** de medidas de similitud geométricas y SIG para darle tratamiento a la componente espacial de los datos en estudios salubristas y una ficha de diagnóstico que permita generar hipótesis o corroborar hallazgos epidemiológicos.
- **Modelación** de la información entorno a la estratificación, los datos geoespaciales y los estudios salubristas.
- **Reutilización de buenas prácticas** relacionadas con estudios estratificados como base para el desarrollo del método y su realización mediante analíticas de datos, que favorezca la incorporación de la espacialidad al sistema de alertas de la vigilancia epidemiológica.

Los enfoques de la propuestas son:

- **Holístico** con el estudio de los indicadores, el espacio en su conjunto y su complejidad, se identifican interacciones, particularidades y procesos que por lo regular no se perciben si se estudian los indicadores por separados y luego se llevan a la cartografía.
- **Estratégico** con la identificación de la situación de los territorios que se estudian y la creación de la ficha de diagnóstico que facilita el establecimiento de objetivos claros a largo plazo y su conjunto de acciones a corto plazo para dar respuesta a las oportunidades y amenazas que impone el entorno, así como las fortalezas y debilidades.

Las cualidades que distinguen al método:

- **Integración**: el método integra indicadores de salud y medidas de similitud geométricas en la estratificación de territorios para facilitar la incorporación del espacio en estudios salubristas. También se distingue por la integración de técnicas de análisis de datos geoespaciales en una solución informática que sirve de soporte tecnológico.
- **Usabilidad**: el enfoque de guía para la estratificación y la interfaz de la instanciación facilitan la integración de la cartografía e indicadores en los estudios sin necesidad de mucho dominio en este campo.
- **Fiabilidad**: la información que brinda se corresponde con el análisis de los indicadores aportados.
- **Flexibilidad**: a partir del uso de indicadores de naturaleza variada y un marco de trabajo para la estratificación se facilita adaptarse a cambios que se deseen incluir en los estudios.

Las premisas:

- Voluntad de las organizaciones y entidades administrativas de la salud pública para su utilización a diferentes niveles administrativos.
- Personal calificado para aplicarlo con rigor científico.
- Disponibilidad de la base cartográfica con la división político administrativa y la información asociada a los indicadores que se incluirán en el estudio.

La aplicabilidad del método se basa en la concepción de que puede ser empleado a diferentes niveles dirección territorial y de administración de la salud. Puede ser extensible a estudios estratificados en otros contextos donde se tengan identificados indicadores y las entidades administrativas asociadas. A tales efectos se debe evaluar la pertinencia de la evaluación de riesgo asociada a los indicadores que se proponen en este método.
Las entradas del método son:

- **Base cartográfica** que está formada por capas vectoriales, de las cuales al menos una debe ser de polígonos y es la capa base que se utilizará como definición de las entidades administrativas a utilizar en la estratificación. El resto de las capas pueden ser de puntos, líneas o polígonos y responden a indicadores geoespaciales.
- **Fuente estadística** cuenta con los indicadores estadísticos a utilizar en la estratificación y sus mediciones se deben corresponder con las entidades geoespaciales definidas por la capa base para la estratificación.

Las salidas del método sirven de soporte para la toma de decisiones, aportan elementos asociados a distribuciones y procesos espaciales útiles para la definición de objetivos y planes en el tratamiento a problemas de salud. Las salidas del método son:

- **Ficha de diagnóstico** contiene los elementos identificados en la estratificación, constituye una base para la toma de decisiones. En esta se detallan los estratos y territorios más afectados, así como los posibles factores asociados. También dispone de un mapa temático a través del cual se puede interactuar con los estratos y territorios.
- **Mapa temático** contiene los estratos clasificados a través de una escala de colores que responde al valor del riesgo obtenido.

La figura 2.1 muestra una representación del método propuesto y sus características son descritas en los siguientes epígrafes del presente capítulo.

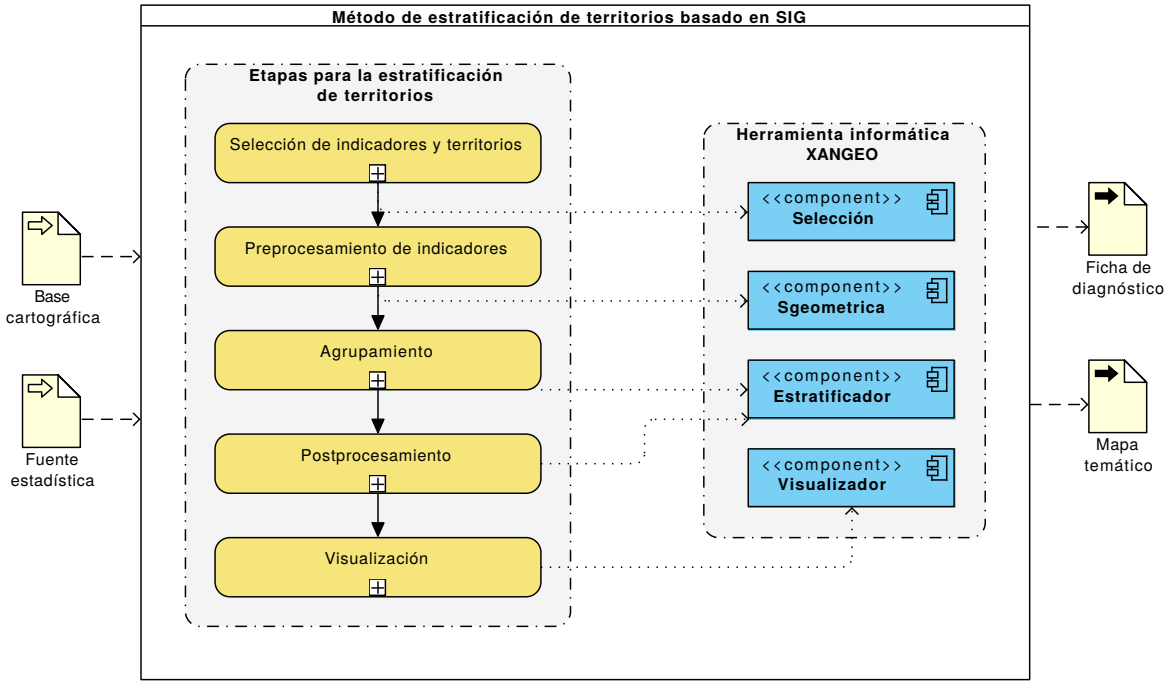


Figura 2.1: *Método de estratificación de territorio basado en SIG y medidas de similitud geométrica, elaboración propia.*

El método de estratificación propuesto está compuesto por un conjunto de etapas con sus procedimientos y el sistema informático XANGEO que consta de los componentes necesarios para la ejecución del mencionado método. Posee un enfoque diferente a los estudios reportados en la literatura, al proponer la integración de indicadores geoespaciales y estadísticos en el estudio. También se integran criterios geométricos para la similitud con el objetivo de lograr grupos más compactos en correspondencia con la primera ley de la geografía.

El método que se propone abarca los procedimientos y etapas de los estudios reportados sobre estratificación, desde un enfoque de la minería de datos geoespaciales. En ella se incluyen cinco etapas destinadas a la selección de los indicadores y la capa base para el estudio, el preprocesamiento de indicadores, el agrupamiento de los territorios, el postprocesamiento para identificar los territorios y estratos más afectados y la visualización de los resultados del estudio. Todas estas etapas serán formalizadas en el epígrafe 2.3.

El sistema informático XANGEO constituye una instanciación del método propuesto e incluye la implementación de las técnicas de minería de datos geoespaciales identificadas en el capítulo 1. También propone un marco de integración entre el SIG QGIS y las principales bibliotecas geoespaciales para su utilización en estudios estratificados. Su arquitectura, principales funcionalidades y componentes son descritos en el epígrafe 2.4.

2.3. Descripción de las etapas que conforman el Método de estratificación de territorios

La figura 2.2 representa la descripción de las etapas que conforman el método para la estratificación de territorios, inicia con la determinación del problema a estudiar, la identificación de los indicadores y la base cartográfica disponible para llevar a cabo el estudio. Incluye técnicas de minería de datos, esencialmente de preprocesamiento y agrupamiento geoespacial como procedimiento para la definición de los estratos.

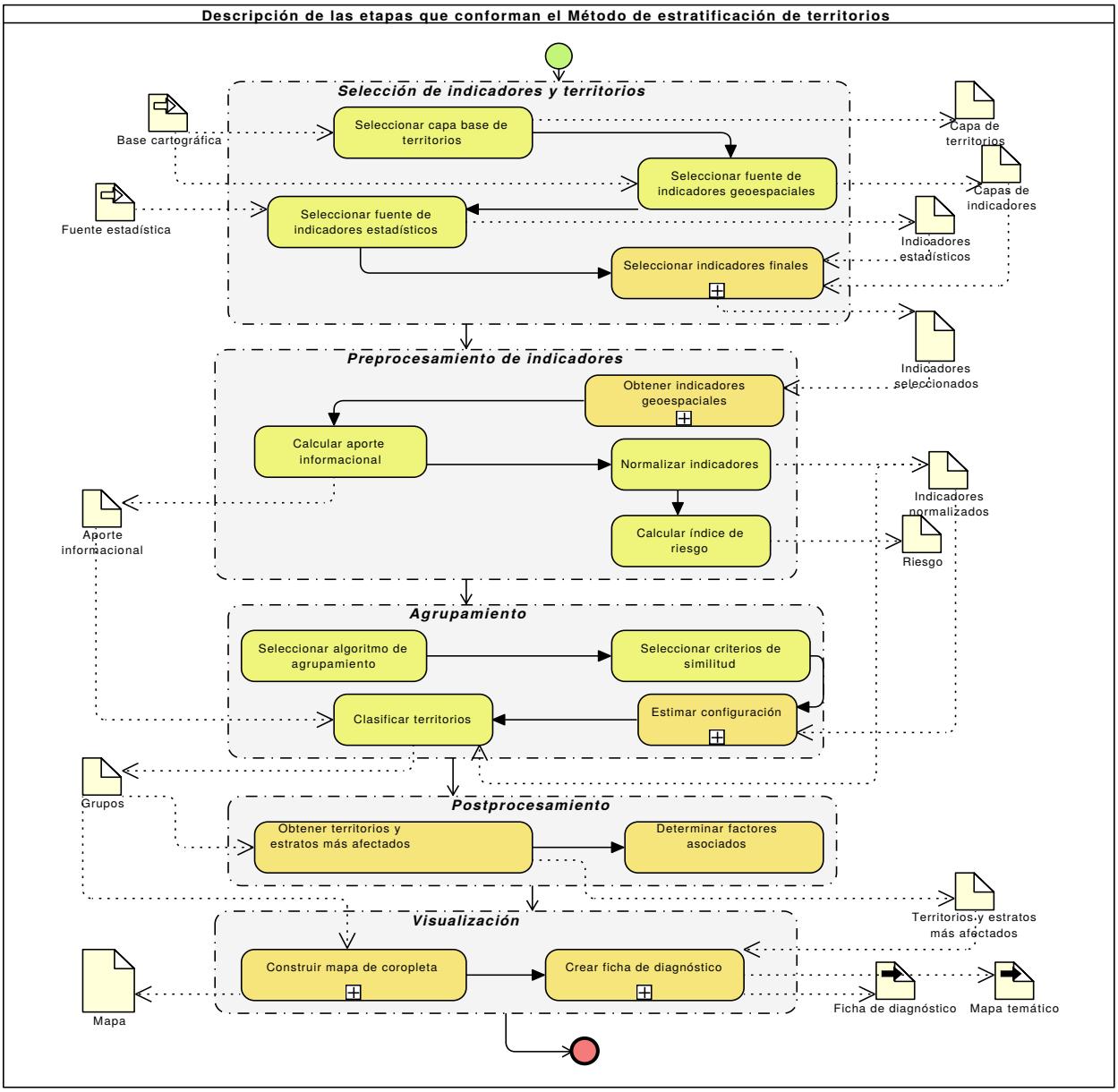


Figura 2.2: *Estrategia para la estratificación de territorios, elaboración propia.*

Con la utilización de este método es posible identificar los territorios y estratos más afectados, de conjunto con los posibles factores asociados, que permite a los salubristas la selección de intervenciones y adecuación de los servicios. También permite la evaluación de los resultados de las intervenciones realizadas, el monitoreo y ajuste de indicadores de acuerdo a los problemas detectados.

2.3.1. Selección de indicadores y territorios

Esta etapa tienen como objetivo elegir los indicadores que se utilizarán en la estratificación y los territorios que serán objeto de estudio, por lo que es necesario disponer de un mapa vectorial e indicadores disponibles en fuentes estadísticos, o recopilados por el investigador utilizando diferentes técnicas y herramientas. Los indicadores pueden ser de naturaleza espacial o temática. Los datos espaciales provienen de la geografía del terreno, y los datos temáticos corresponden a la información de otras fuentes como son datos de la población, factores de riesgos e indicadores de salud, por solo mencionar algunos. La representación de esta etapa la puede consultar en la figura 2.3.

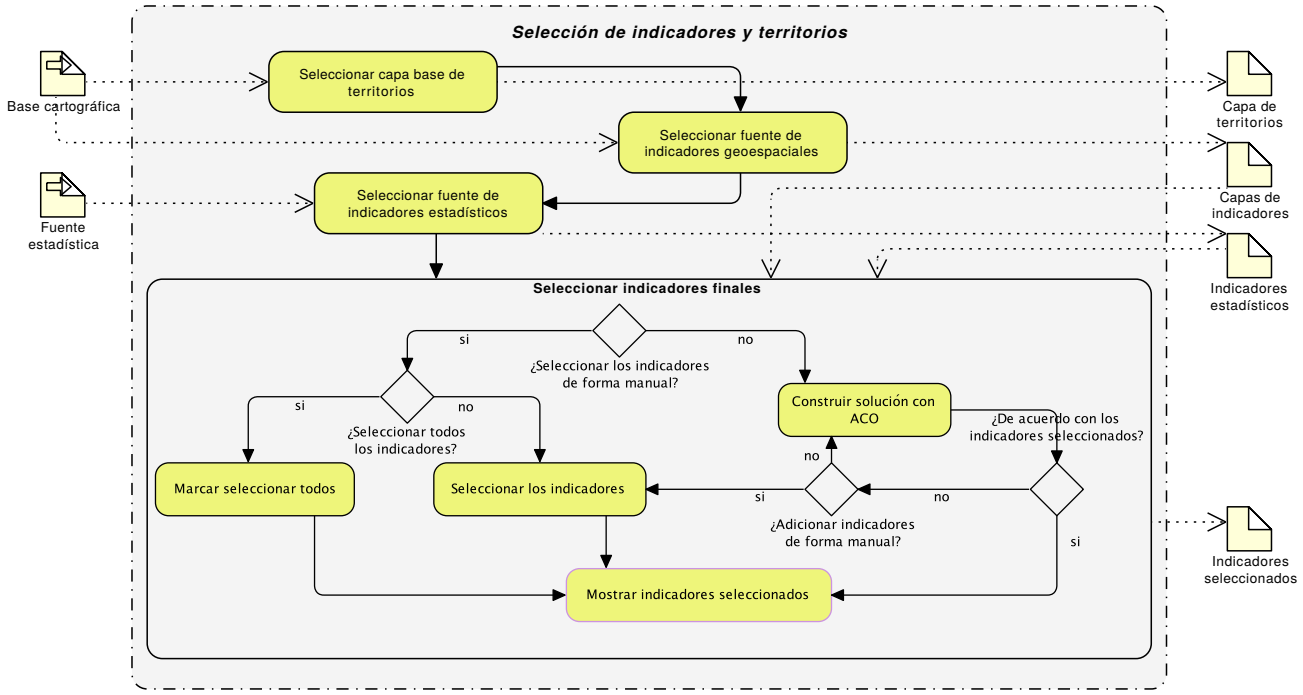


Figura 2.3: *Selección de indicadores y territorios, elaboración propia.*

La etapa inicia con la selección de la capa de polígonos que define los objetos geoespaciales que representan a cada territorio objeto de estudio. Desde esta capa se extrae la información necesaria, identificadores o nombres de los territorios, que permita asociar el objeto geoespacial con los valores de las fuentes estadísticas. Posteriormente se seleccionan los indicadores geoespaciales en correspondencia con mediciones de variables en el espacio y por

tanto se encuentran georreferenciadas. Estos indicadores se representan mediante capas vectoriales y en la etapa de Preprocesamiento de indicadores se les proporciona el tratamiento para la integración en el estudio. La selección de indicadores estadísticos parte de la identificación de los indicadores disponibles en la fuente estadística. En el paso posterior se puede decidir si se escogen los indicadores de forma manual o se utiliza alguna de las técnicas automáticas. Los indicadores se escogen de forma manual cuando el investigador tiene identificado cuales son los que inciden en el problema objeto de estudio o cuando realiza una evaluación de los resultados de las intervenciones realizadas a partir de estudios realizados, e incluye el monitoreo y ajuste de indicadores de acuerdo a los problemas detectados.

En la selección de atributos de forma automática están relacionadas distintas áreas como la del reconocimiento de patrones [179], el aprendizaje automático [180] y la minería de datos [181], suele estar presente en las etapas previas de la minería de datos, ya sean supervisadas o no. Se encuentran diversas definiciones que la autora resume en: encontrar el subconjunto de atributos del conjunto de datos original, que mejor describa los objetos del dominio. En la selección de atributos de forma automática se reduce la dimensionalidad del conjunto, a través de la selección del subconjunto de mejor desempeño bajo algún criterio de clasificación. La selección de atributos reporta beneficios al eliminar los rasgos irrelevantes y redundantes [182], lo que posibilita una mejor representación de la información original y una disminución del costo computacional.

Los procedimientos de selección de atributos requieren un método de generación de subconjuntos (basado en un proceso de búsqueda). En la presente investigación se utiliza la metaheurística basada en colonia de hormigas (ACO). Las hormigas construyen las soluciones de manera probabilística, guiándose por un rastro de feromona artificial y por una información calculada a priori de manera heurística [183, 184, 185]. Se obtienen las características más visitadas, es decir el subconjunto de características que mejor describe a los objetos.

Al finalizar el investigador tiene la opción de incluir los indicadores que estime necesarios para su estudio y que no se encuentren en la solución encontrada por el procedimiento de selección. El resultado de esta etapa es un conjunto de indicadores, tanto estadísticos como geoespaciales, que se utilizarán para realizar la estratificación. En la siguiente etapa estos indicadores son preprocesados para facilitar el análisis y la interpretación de los resultados.

2.3.2. Preprocesamiento

Esta etapa tiene como objetivo preparar la base cartográfica y los indicadores seleccionados para realizar la estratificación. Se obtienen los valores de los indicadores geoespaciales a partir de las ecuaciones 2.1, 2.2 y 2.3. Posteriormente se calcula el aporte informacional y se normalizan los valores de los indicadores seleccionados. La base cartográfica debe estar formada por n capas con $n \geq 1$, para el caso donde $n = 1$ entonces, la capa debe ser de polígonos y contiene los objetos geoespaciales asociados a los territorios objetos de estudio. La representación de esta etapa la puede consultar en la figura 2.4.

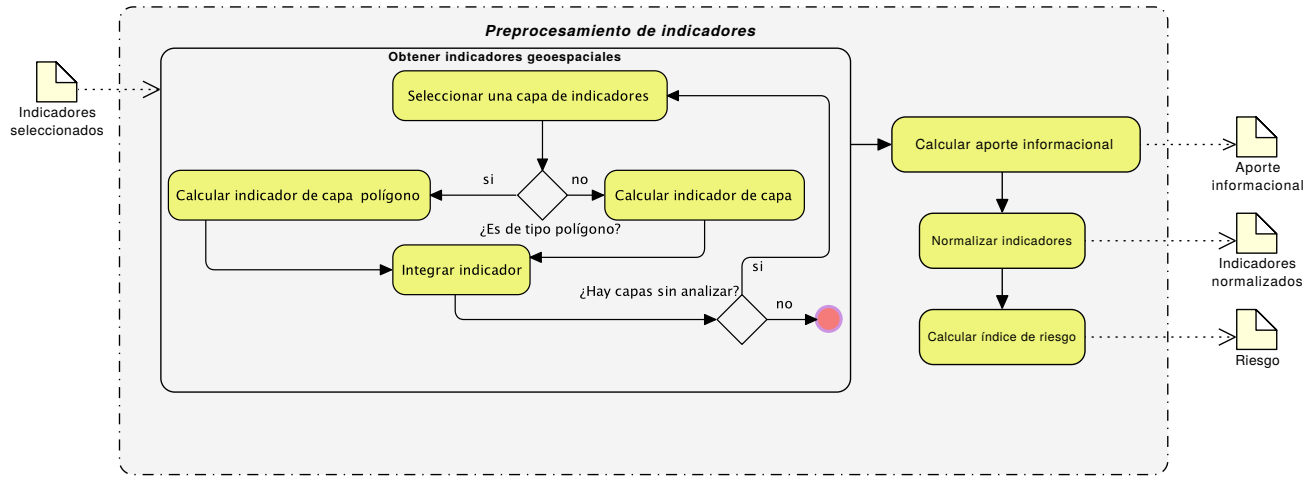


Figura 2.4: *Etapa de preprocesamiento, elaboración propia.*

2.3.2.1. Obtener indicadores geoespaciales

Los indicadores para esta etapa se clasifican según su naturaleza geoespacial en: indicadores geoespaciales e indicadores temáticos. Los indicadores geoespaciales se obtienen a partir de capas de la base cartográfica. Cada capa representa un indicador y en correspondencia con el tipo de objeto geoespacial de la capa se obtiene el indicador de capa. Para obtener el indicador de capa de tipo geoespacial: punto, línea o polígono se definen las ecuaciones 2.1, 2.2 y 2.3 respectivamente.

$$I_{ij} = \frac{n}{A(p_i)}, \forall p_i \in P \tag{2.1}$$

Donde P es el conjunto de polígonos que representan a los territorios, p_i se corresponde con un polígono dentro del conjunto P, n es la cantidad de puntos que se encuentran dentro del polígono p_i y $A(p_i)$ constituye el área del polígono p_i. De esta forma se obtiene el indicador de capa y se corresponde con la densidad de puntos dentro del polígono p_i.

$$I_{ij} = \frac{n}{A(p_i)}, \forall p_i \in P \tag{2.2}$$

Para esta ecuación p_i, P y $A(p_i)$ se corresponden con la descripción anterior, n representa la cantidad de líneas contenidas dentro de p_i.

$$I_{ij} = \frac{\sum_{k=1}^{n} A(p_{i,k})}{A(p_i)}, \forall p_i \in P \tag{2.3}$$

Donde n representa la cantidad de polígonos contenidos por el polígono p_i y $p_{i,k}$ representa a cada uno de los polígonos contenidos en p_i.

Por cada indicador obtenido a partir de los indicadores geoespaciales y por cada territorio se asocia el valor calculado para el indicador en Integrar indicador. Al terminar este procedimiento todos los indicadores poseen valores numéricos y se encuentran listos para las tareas posteriores.

2.3.2.2. Calcular aporte informacional de los indicadores

El aporte informacional de los indicadores (en función del dominio de cada indicador) se obtiene mediante el coeficiente de variación a partir de la ecuación:

$$\sigma^2 = \sqrt{\frac{1}{t}\sum_{i=1}^{t}(V_i - \bar{x})^2} \tag{2.4}$$

Donde t es el total de la población finita de los datos, $\bar{x}$ media de los datos, V_i valor i, donde $i = 1, 2, ..., t$. Posteriormente se calcula la desviación estándar σ como la raíz cuadrada de la varianza de los datos. El coeficiente de variación (2.4) obtiene la dispersión de los datos en función de su media y se determina:

$$w_k = \frac{\sigma}{\bar{x}} \tag{2.5}$$

Donde: σ: desviación estándar y $\bar{x}$: media. Finalmente w_k es aporte informacional del indicador k, para $k = 1, 2, ..., t$.

2.3.2.3. Normalizar indicadores

La normalización de los datos es una técnica que se aplica a un conjunto de datos para reducir su redundancia. A partir de los diferentes dominios en los que se presentan los valores de estos indicadores, se normalizan para evitar que atributos con valores muy altos tengan mayor peso que atributos con valores bajos. También se propone determinar el índice de riesgo de los territorios con un valor numérico en el intervalo [0,1].

Desde la estadística y en las ciencias de los datos se han introducido diferentes estrategias para la normalización. En esta investigación se propone utilizar la ecuación denominada característica escala o normalización basada en la unidad. Desde este enfoque todos los valores se transforman en el rango de [0,1].

$$X_{i,f} = \frac{X_{i,f} - f_{min}}{f_{max} - f_{min}} \tag{2.6}$$

En esta ecuación $X_{i,f}$ es el valor i del atributo f, con $i = 1, 2, ..., t; f_{min}, f_{max}$ mínimo y máximo valor del indicador respectivamente.

Es importante tener en cuenta que cuando los datos presentan ruido esta técnica de normalización puede no ser efectiva y se recomienda primeramente dar tratamiento al ruido a partir de las técnicas disponibles.

2.3.2.4. Calcular índice de riesgo

El índice de riesgo asociado a cada indicador se utiliza para determinar el riesgo asociado a cada territorio. Para determinar el índice de riesgo, los indicadores se clasifican según el impacto que tiene su valor. Los que a mayor valor del indicador aportan más al riesgo, ecuación 2.7 y los que a menor valor aportan más al índice de riesgo, ecuación 2.8.

$$\tau_{x_{i,k}} = \frac{x_{i,k}}{k_{max}} \tag{2.7}$$

$$\tau_{x_{i,k}} = 1 - \frac{x_{i,k}}{k_{max}} \tag{2.8}$$

Donde τ representa la aportación de riesgo, $x_{i,k}$: Es el valor que tiene el territorio i para el indicador k y k_{max} es el mayor valor que toma el indicador k.

Para la determinación del riesgo de un territorio se utiliza la ecuación 2.9, que integra el aporte informacional de cada indicador con su índice de riesgo.

$$\gamma(x_m) = \frac{1}{\sum_{i=1}^{n} w_k} \sum_{i=1}^{n} w_k \tau_i \tag{2.9}$$

A partir de este valor de riesgo en etapas posteriores se guiarán los análisis de la dependencia espacial y la evaluación del clasificador para la tarea de estratificación.

2.3.3. Agrupamiento

Esta etapa tiene como objetivo la construcción de los estratos. Para ello se clasifican los territorios en grupos homogéneos (estratos), utilizando las técnicas y los algoritmos de agrupamiento descritos en los epígrafes 1.2 y 1.4. Se inicia con la selección del algoritmo de agrupamiento a partir de las características del estudio que se realiza y de la experiencia del investigador o del equipo que ejecuta la investigación. La representación de esta etapa la puede consultar en la figura 2.5.

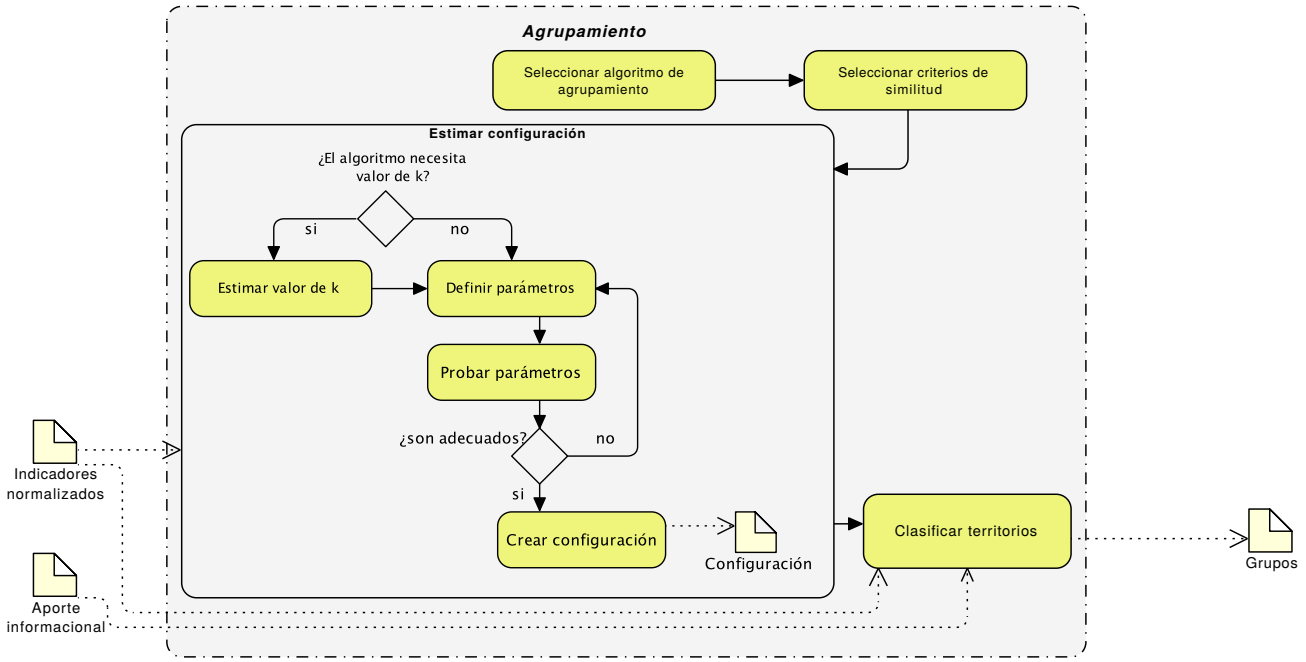


Figura 2.5: *Etapa de agrupamiento, elaboración propia.*

2.3.3.1. Seleccionar criterios de similitud

Los criterios de similitud deben seleccionarse a partir del objetivo que persigue la estratificación y se incorporan criterios geométricos a partir de los identificados en el capítulo 1. Estos criterios y su elección constituyen un paso relevante para la estratificación desde el enfoque que se propone en este trabajo.

Algoritmo 1 Distancia basada en la ubicación geoespacial de los polígonos

Entrada: Una capa C de polígonos, identificadores de los dos polígonos p_i, p_j

Salida: distancia d_{p_i,p_j}

1: $territorio_i = C.getFeature(p_i)$
2: $territorio_j = C.getFeature(p_j)$
3: **Si** $geometry(territorio_i).isMultipart() \parallel geometry(territorio_j).isMultipart()$ **entonces**
4: $point_i = centroid(territorio_i).asPoint()$

Para calcular las coordenadas de punto correspondiente al centroide:

$$\begin{aligned} c_x &= \frac{1}{6A}\sum_{i=0}^{N-1}(x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i) \\ c_y &= \frac{1}{6A}\sum_{i=0}^{N-1}(y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i) \end{aligned} \tag{2.10}$$

5: $point_j = centroid(territorio_j).asPoint()$
6: $s_{p_i,p_j} = measureLine(point_i, point_j)$
7: **Sino**
8: $d_{p_i,p_j} = QgsGeometry.distance(point_i, point_j)$
9: **Fin Si**
10: **Retornar** d_{p_i,p_j}

Para determinar la similitud de dos objetos en función de su conectividad se obtiene la longitud de los lados comunes, por la tesis de que si tienen al menos un lado en común están conectados, de esta manera objetos vecinos son más similares [53]:

$$\delta(P_i, P_j) = 1 - \frac{min(p_{P_i}, p_{P_j}) - x}{max(p_{P_i}, p_{P_j}) - x} \tag{2.11}$$

Donde P es el perímetro del polígono y x es la longitud de los lados en común en los polígonos Pi y Pj.

Algoritmo 2 Distancia basada en la conectividad de los polígonos

Entrada: Una capa C de polígonos, identificadores de los dos polígonos p_i, p_j

Salida: distancia d_{p_i,p_j}

1: $territorio_i = C.getFeature(p_i)$
2: $territorio_j = C.getFeature(p_j)$
3: **Si** $intersects(territorio_i, territorio_j)$ **entonces**
4: $line = intersection(territorio_i, territorio_j)$
5: $perim_i = measurePerimeter(territorio_i)$
6: $perim_j = measurePerimeter(territorio_j)$
7: $conect = measureLine(line)$
8: $d_{p_i,p_j} = \frac{\text{mín}(perim_i, perim_j) - conect}{\text{máx}(perim_i, perim_j) - conect}$
9: **Fin Si**
10: **Retornar** d_{p_i,p_j}

La similitud de un polígono está dada en la proximidad a uno del resultado de la división del más grande por el más pequeño, ya que una división indica la cantidad de veces que cabe el divisor dentro del dividendo. Para esto primero se determina una forma de calcular el área en polígonos irregulares, para después establecer la relación planteada entre sus áreas. La similitud según el criterio del tamaño se determina a partir de la ecuación:

$$\delta(P_i, P_j) = 1 - \frac{min(A_{P_i}, A_{P_j})}{max(A_{P_i}, A_{P_j})} \tag{2.12}$$

Donde A representa el área de los polígonos.

Algoritmo 3 Distancia basada en el tamaño de los polígonos

Entrada: Una capa C de polígonos, identificadores de los dos polígonos p_i, p_j

Salida: distancia d_{p_i,p_j}

1: $territorio_i = C.getFeature(p_i)$
2: $territorio_j = C.getFeature(p_j)$
3: $area_i = measureArea(territorio_i)$
4: $area_j = measureArea(territorio_j)$
5: $d_{p_i,p_j} = 1 - \frac{\text{mín}(area_i, area_j)}{\text{máx}(area_i, area_j)}$
6: **Retornar** d_{p_i,p_j}

Se debe significar que la utilización de un criterio o combinaciones de ellos dependen en gran medida del dominio de las características del problema que se estudia y sus resultados no se contradicen, solo muestran alternativas de análisis y puntos de vistas. Tampoco se contradicen con la utilización de las ecuaciones descritas en el epígrafe 1.5.

2.3.3.2. Estimar configuración y clasificar los territorios

La configuración a los efectos de esta investigación se corresponde con los parámetros que necesita el algoritmo de agrupamiento seleccionado para la construcción de los grupos. Para los algoritmos que necesitan la cantidad de grupos, en la presente investigación se estima utilizando el coeficiente de silueta a partir de los resultados mostrados en otras investigaciones [186, 20].

El coeficiente de silueta es una métrica para evaluar la heterogeneidad de los grupos producidos por algoritmos de aprendizaje no supervisado. Un valor más alto de este índice indica un caso más deseable del número de clústeres. El coeficiente de silueta para una observación i se denota como $s(i)$ y se define como:

$$s(i) = \frac{b - a}{max(a, b)} \tag{2.13}$$

Donde: a es el promedio de las disimilitudes (o distancias) de la observación i con las demás observaciones del clúster al que pertenece i. Y b es la distancia mínima a otro clúster diferente al que contiene a la observación i. Ese clúster es la segunda mejor opción para i y se lo denomina vecindad de i. El valor de $s(i)$ puede ser obtenido

combinando los valores de a y b como se muestra a continuación:

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & si \quad a(i) < b(i) \\ 0 & si \quad a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1 & si \quad a(i) > b(i) \end{cases} \tag{2.14}$$

El coeficiente de silueta es un valor comprendido entre -1 y 1, $-1 \leq s(i) \leq 1$. Al analizar las posibles soluciones para que el coeficiente de silueta sea cercano a uno, el valor de b tiene que ser mayor al de a. Esto significa que la distancia de la observación i a los clústeres vecinos es suficientemente grande para que su pertenencia al clúster actual sea la correcta. Es decir, no es similar a sus vecinos. Un valor de $s(i)$ que sea cercano a cero nos va a indicar que la observación i está en la frontera de dos clústeres y si el valor de $s(i)$ es negativo, entonces la observación i debería ser asignada al clúster más cercano como se muestra a continuación:

- $s(i) \approx 1$, la observación i está bien asignada a su *clúster*.
- $s(i) \approx 0$, la observación i está entre dos *clúster*.
- $s(i) \approx -1$, la observación i está mal asignada a su *clúster*.

El coeficiente de silueta se calcula como el promedio de todos los $s(i)$ para todas las observaciones del conjunto de datos.

Los parámetros necesarios para otros algoritmos de agrupamientos o que necesiten la cantidad de grupos estimada como se explicó anteriormente son definidos por el investigador y se prueban hasta tanto no se obtengan valores adecuados para el mismo. Este paso depende en gran medida de la experiencia del investigador.

Una vez probados los parámetros y creada la configuración del clasificador, se procede a clasificar los territorios en grupos. Como resultado de esta etapa se obtienen los grupos o estratos con los valores de riesgo asociados a cada grupo.

2.3.4. Postprocesamiento

Esta etapa tiene como objetivo la obtención de los territorios y estratos más afectados por indicadores. La representación de esta etapa la puede consultar en la figura 2.6. Se debe especificar que la etapa Postprocesamiento en la estratificación de territorios recibe los datos resultantes del Agrupamiento. Estos datos son utilizados para mitigar los riesgos y las estrategias implican asignaciones de recursos, análisis que pueden ser epidemiológicos y sociales. Hay evidencia en la literatura de utilización de estos resultados para análisis de percepción o de sentimientos en redes sociales en estudios epidemiológicos [187]. Este método de evaluación es mucho más rápido y menos costoso que las tradicionales encuestas.

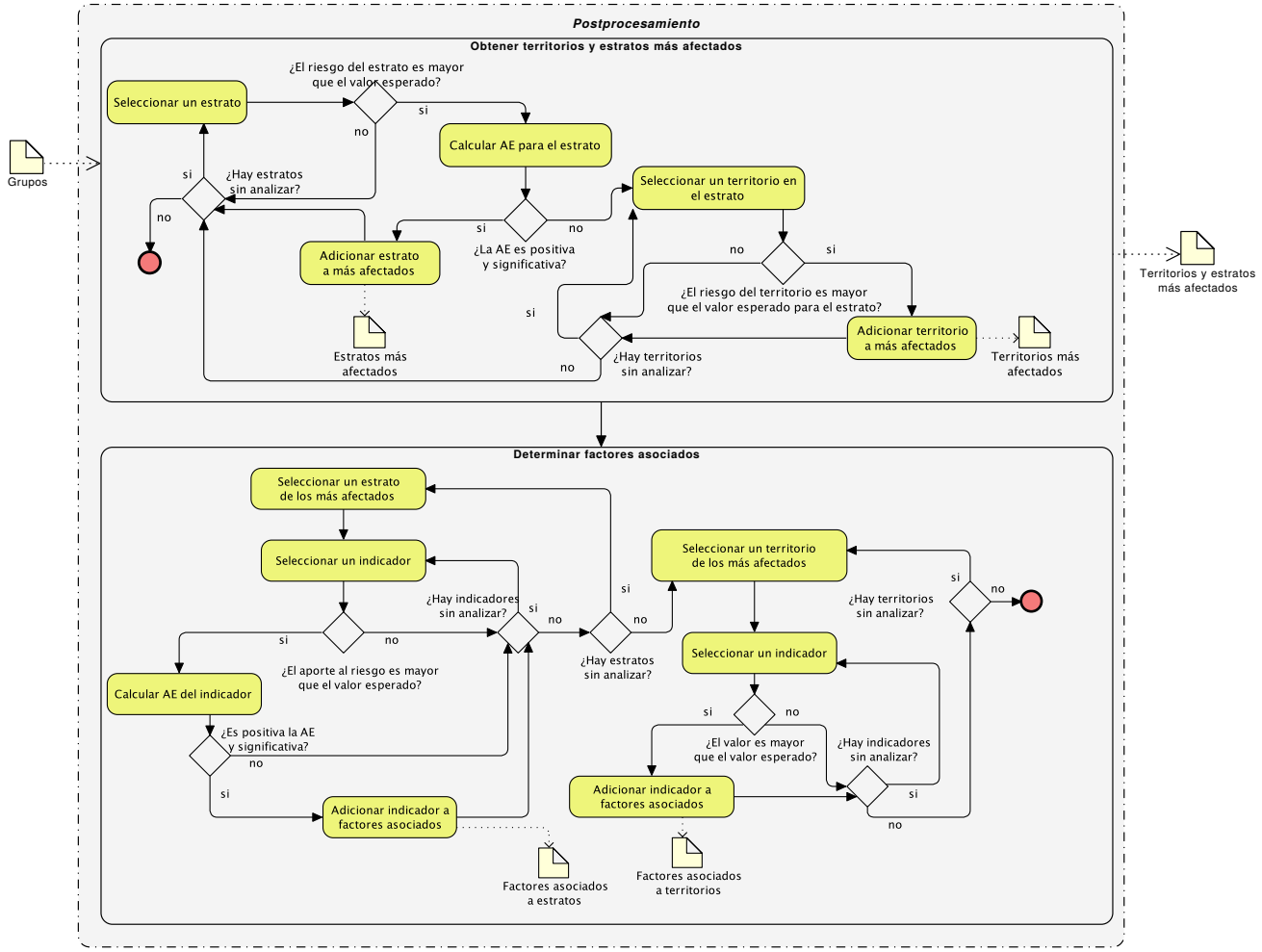


Figura 2.6: *Etapa de Postprocesamiento, elaboración propia.*

2.3.4.1. Obtener territorios y estratos más afectados

Los estratos y territorios más afectados son los que tienen mayor aporte al riesgo, indicadores del estrato con Autocorrelación espacial (AE) positiva y esta es significativa y superior al valor esperado. La dependencia o autocorrelación espacial aparece como consecuencia de la existencia de una relación funcional entre lo que ocurre en un punto determinado del espacio y lo que ocurre en otro lugar [49]. La autocorrelación espacial se define como la medida cuantitativa de la concentración o dispersión de los valores de una variable en un mapa. Con ella es posible determinar el grado en que los datos espaciales de una unidad geográfica son similares a otros en unidades geográficas próximas [188]. El concepto de autocorrelación espacial parte del principio de Tobler o primera ley de la geografía que refiere que todo está relacionado, pero objetos más cercanos tienen más relación.

Por lo tanto, se intenta medir la correlación que el riesgo obtenido en las etapas anteriores tiene en diferentes unidades espaciales contiguas en una perspectiva horizontal dando lugar a una de estas tres posibilidades [188]:

- Autocorrelación espacial positiva: las unidades espaciales vecinas presentan valores próximos. Indica una tendencia al agrupamiento de las unidades espaciales.
- Autocorrelación espacial negativa: las unidades espaciales vecinas presentan valores muy disímiles. Indica una tendencia a la dispersión de las unidades espaciales.
- Sin autocorrelación: no ocurre ninguna de las dos situaciones anteriores. Por lo tanto, los valores de las unidades espaciales vecinas presentan valores producidos en forma aleatoria.

La autocorrelación espacial puede ser univariada o bivariada. En un gráfico de dispersión, en el eje x aparecen los valores estandarizados de una variable para cada unidad espacial y en el eje y se encuentran los valores estandarizados del promedio de los valores de las unidades espaciales vecinas para la misma variable (en el caso de la autocorrelación espacial univariada) o de una segunda variable (autocorrelación espacial bivariada). En ambos casos, la recta de regresión lineal muestra el grado de asociación entre la variable y los valores contiguos de la misma u otra variable considerada.

Para detectar y medir la AE, a través del Índice de Moran los resultados varían entre el –1 y 1, representando las correlaciones mínimas (máxima dispersión) y máximas (máxima concentración) respectivamente y el cero significa un patrón espacial totalmente aleatorio.

Para definir si una autocorrelación espacial es significativa se realiza una prueba de hipótesis y así se comprueba si la configuración espacial de la variable se produce aleatoriamente, es decir si se cumplen o no los supuestos del modelo a partir de estimar si un estadístico muestral difiere significativamente de lo esperado aleatoriamente. Esta prueba se efectúa al ubicar el coeficiente de Moran dentro de una curva normal de probabilidades.

Al realizar este tipo de test en el campo del análisis geoespacial, inicialmente se define la hipótesis nula que responde a la afirmación de que la configuración espacial se produce de manera aleatoria y la alternativa configuración espacial no se produce de manera aleatoria. Luego se especifica el nivel de significación que indica la probabilidad de rechazar la hipótesis nula siendo ésta verdadera. Por lo tanto, es la mayor probabilidad que se está dispuesto a arriesgar a cometer un error de decisión de aceptar la hipótesis alternativa. Se suele elegir de acuerdo a la importancia del problema y generalmente es del cinco por ciento (0.05) y el uno por ciento (0.01). Asimismo, el p-valor es el resultado que nos brinda la prueba de hipótesis. Si el nivel de significación es superior al p-valor, se rechaza la hipótesis nula y se acepta la alternativa. Por el contrario, se comprueba la hipótesis nula, es decir, que la configuración espacial se produce de forma aleatoria.

A partir de este procedimiento se identifican los estratos más afectados que se corresponden con los que tienen valores de riesgo superiores al valor esperado y que a lo interno poseen una autocorrelación positiva y significativa. Los territorios más afectados son el resultado del análisis de aquellos que tienen un riesgo superior al valor esperado y no se puede descartar una configuración espacial aleatoria.

2.3.4.2. Determinar factores asociados

Para obtener los factores asociados al comportamiento de los estratos y territorios más afectados se sigue el siguiente procedimiento. Primeramente se analizan los estratos más afectados obtenidos por el algoritmo anterior. Luego se examinan todos los indicadores que poseen aporte al riesgo superior al valor esperado y se calcula la AE a lo interno del estrato. Si la AE es positiva y la prueba de hipótesis arroja resultados significativos entonces se incluye el indicador en la lista de factores asociados. De lo contrario no existen evidencias suficientes para incluirlo, aunque no se puede descartar la hipótesis de que esté incidiendo de forma negativa sobre un territorio en particular y debe ser identificado por el siguiente procedimiento.

Por otra parte se analizan los factores que afectan a los territorios de forma independiente. Para este procedimiento se parte del enfoque de que si el riesgo de un indicador es mayor que el valor esperado y tiene una AE positiva y significativa entonces debe ser identificado entre los que afectan a los estratos. En correspondencia con esto, cuando su riesgo es mayor que el valor esperado y sigue una configuración espacial aleatoria entonces este indicador es identificado como un factor que afecta a los territorios para los cuales tiene un riesgo superior al valor esperado.

2.3.5. Visualización

Esta etapa tiene como objetivo representar en un mapa temático del tipo coropleto o coropléticos cada grupo homogéneo de territorios, esencial para comprender las realidades espaciales [189]. La representación de esta etapa la puede consultar en la figura 2.7.

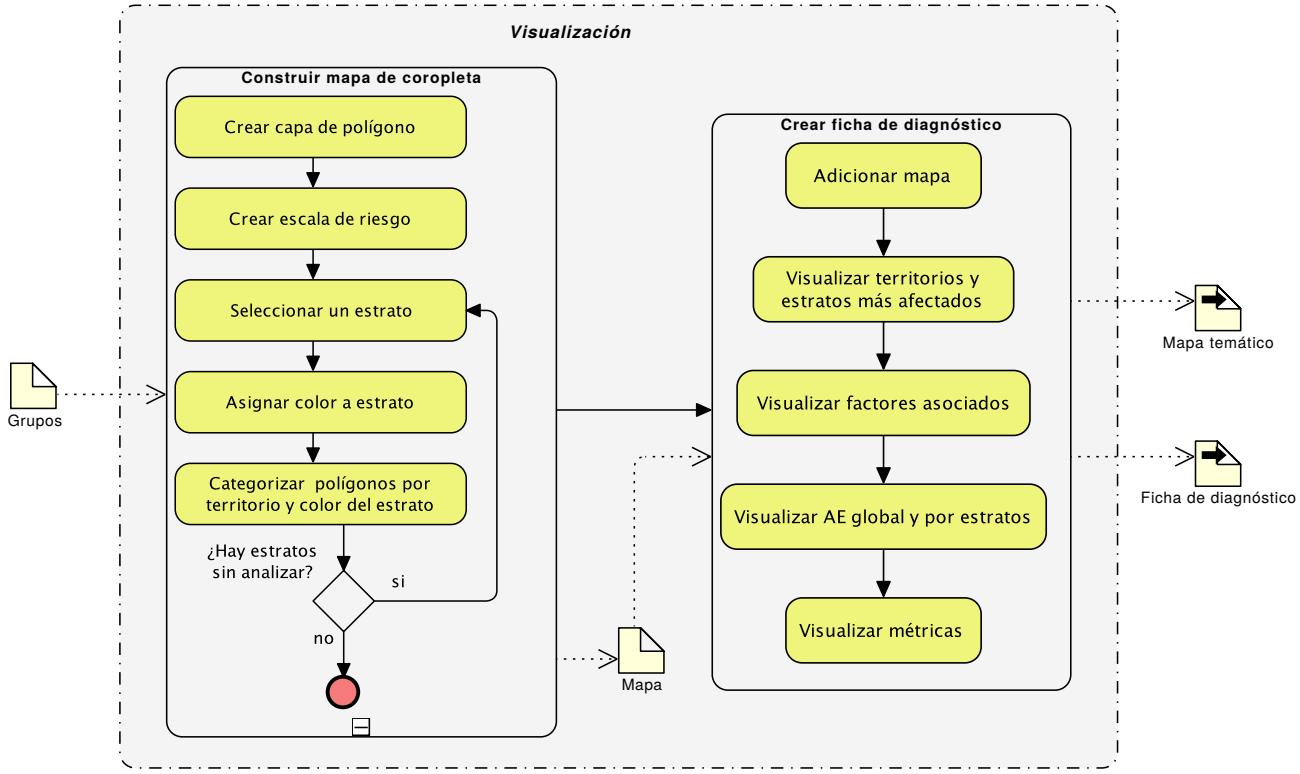


Figura 2.7: *Etapa de Visualización, elaboración propia.*

2.3.5.1. Construir mapa de coropleta

Los mapas temáticos coropletos reportados en la literatura tienen la particularidad de que las regiones se colorean a partir de una medida estadística, como puede ser la densidad de población o el ingreso por habitante. Este tipo de mapa facilita la comparación de una medida estadística de una región con la de otra o muestra la variabilidad de esta para una región dada [190]. Se reporta en fenómenos como la variación espacial que coincide con los límites de las unidades de medición, por ejemplo se distribuye de manera uniforme dentro de cada unidad de medición [191]. En esta investigación cada grupo homogéneo de territorios es representado por un color asignado al estrato en función del riesgo promedio del estrato.

2.3.5.2. Crear ficha de diagnóstico

La ficha de diagnóstico es la herramienta fundamental que brinda el método propuesto como soporte a la toma de decisiones. Este procedimiento inicia con la adición del mapa temático construido anteriormente, en el que se representan los estratos con colores. Estos colores representan el riesgo promedio del riesgo del estrato.
Posteriormente se visualizan en la ficha los territorios y estratos más afectados y sobre ellos se puede identificar los factores asociados. La ficha también permite darle seguimiento al problema que se estudia y por eso se adicionan los resultados del análisis exploratorio de datos, fundamentalmente la autocorrelación espacial y la evaluación de las métricas. Estos elementos permiten que luego de realizado un estudio y diseñada la intervención, se puedan

realizar estudios para evaluar los resultados de la intervención.

2.4. Herramienta informática XANGEO

Esta instanciación tiene como objetivo demostrar la viabilidad del método propuesto y facilitar la evaluación concreta de su idoneidad en la estratificación de territorios. El método propuesto, ha sido implementado como un complemento para el SIG QGIS cuenta con los siguientes requisitos funcionales:

- RF 1: Importar indicadores estadísticos desde una hoja de cálculo.
- RF 2: Obtener características cartográficas a través de QGIS.
- RF 3: Construir estratificación.
- RF 3.1: Construir estratos.
- RF 3.2: Visualizar estratos construidos en mapa temático
- RF 4: Gestionar las estratificaciones.
- RF 4.1: Adicionar estratificación.
- RF 4.2: Mostrar estratificación.
- RF 4.3: Eliminar estratificación.
- RF 5: Exportar mapa temático de una estratificación como imagen.
- RF 6: Exportar estratificación hacia una hoja de cálculo.

La arquitectura de un sistema informático establece entre otros aspectos la base para la evolución y el mantenimiento del software. La separación e independencia son fundamentales para el diseño arquitectónico porque permiten localizar cambios. El diseño del sistema y la organización está regido por un estilo arquitectónico de Arquitectura en capas. Esta posibilita que la funcionalidad del sistema está organizada en capas separadas, y cada una apoya sólo en las facilidades y los servicios ofrecidos por la capa debajo de ella. Este enfoque en capas soporta el desarrollo incremental de sistemas [192], lo cual puede ser observado en la figura 2.8.

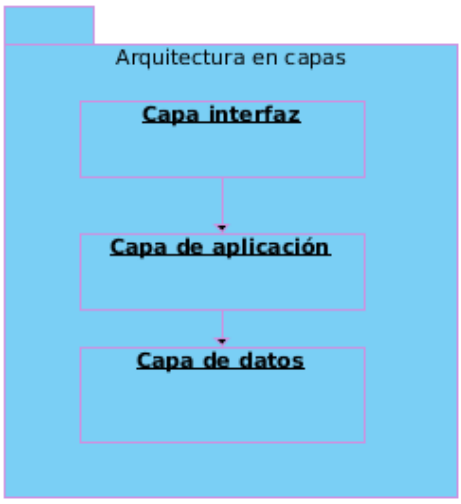


Figura 2.8: *Diagrama Arquitectura en capas, elaboración propia*

La aplicación de este estilo arquitectónico posibilita que las funcionalidades del sistema estén organizadas en capas separadas y cada una se apoya sólo en las facilidades y los servicios ofrecidos por la capa inmediatamente debajo de ella [192]. La capa inicial es responsable de implementar la interfaz de usuario, la segunda capa implementa la funcionalidad del sistema, la capa de base de datos, ofrece administración de transacciones y almacenamiento constante de datos.

Con el objetivo de traducir el modelo del diseño a software operativo [193] se presenta el Diagrama de componentes (figura 2.9) el cual muestra la composición de la estratificación contenida en el complemento XANGEO.

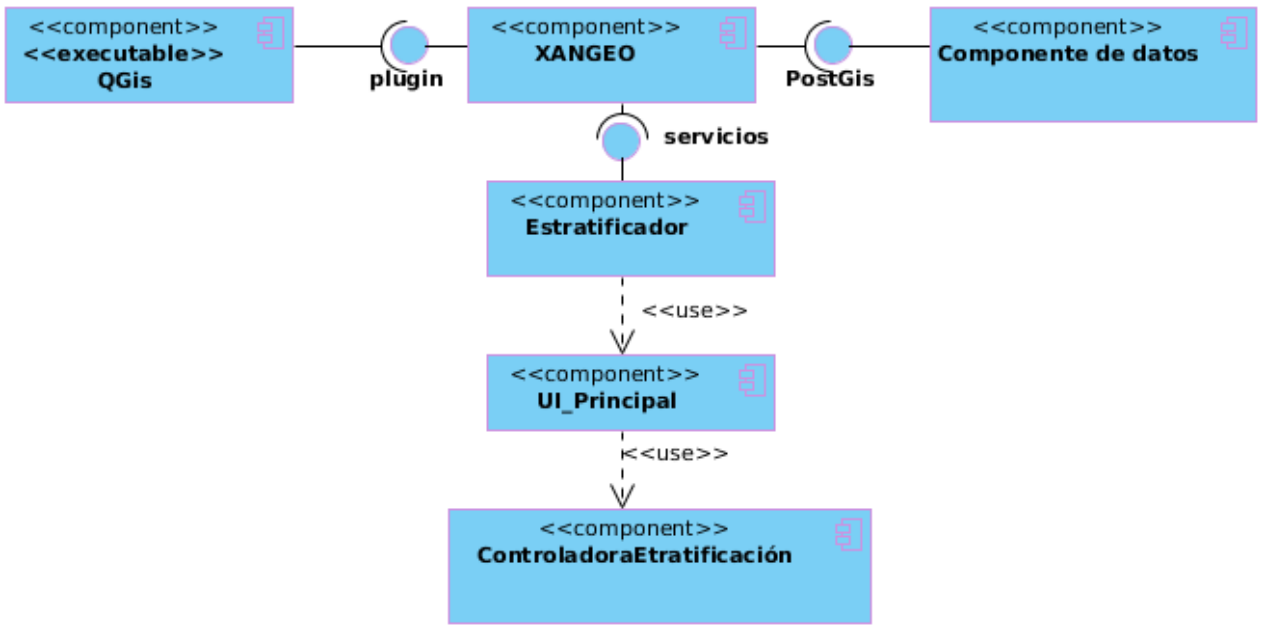


Figura 2.9: *Diagrama de componentes, elaboración propia*

La figura 2.10 muestra la ventana inicial de la solución obtenida para la instanciación del método. Desde esta interfaz accede a los elementos de configuración de la estratificación, se establece la capa base y los indicadores que se estudian.

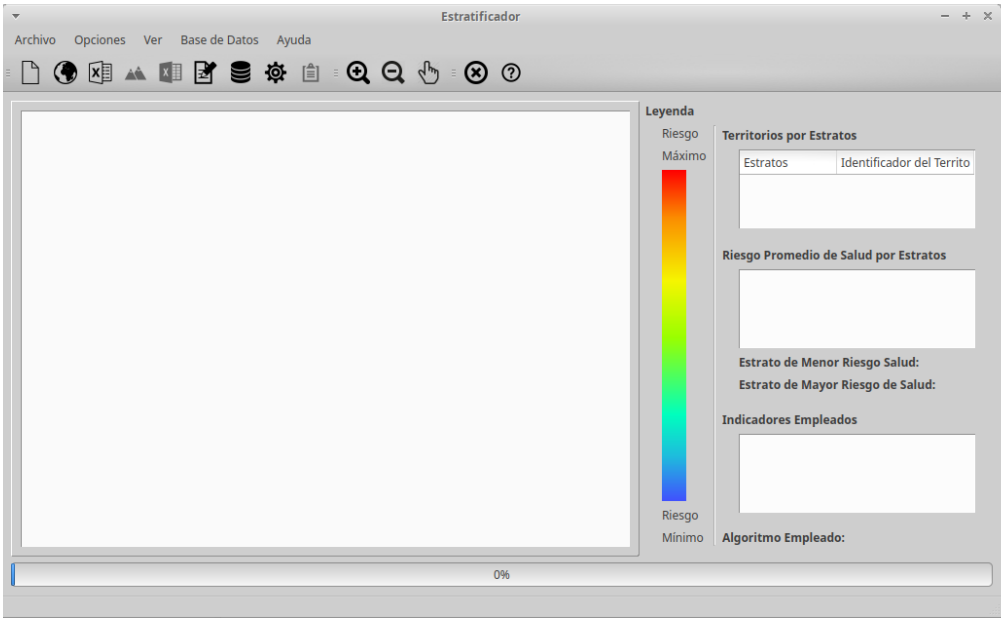


Figura 2.10: *Vista principal, tomada de XANGEO*

2.5. Conclusiones del capítulo

Durante este capítulo se presentó un método de estratificación de territorios basado en SIG y medidas de similitud geométrica que combina algoritmos y descripciones textuales identificados en enfoques precedentes e integra criterios de similitud geométricos con el objetivo de obtener estratos más compactos. A partir de la ejecución de las etapas previstas según el paradigma empleado se arrojan las siguientes conclusiones:

- La identificación de los constructos dentro del objeto de estudio que se aborda en la presente investigación facilitó la descripción adecuada del problema y su solución a partir del método propuesto, que además integra los enfoques aportados en investigaciones precedentes con relación a estudios estratificados y la componente espacial de los datos en el espacio de solución del problema.
- La integración de medidas de similitud geométricas en el proceso de estratificación de territorios facilita la incorporación del espacio en estudios salubristas y constituye una alternativa de análisis alineada al principio de la primera ley de la geografía.
- La instanciación del método como un componente para el Sistema de Información Geográfica QGIS facilita la evaluación de la viabilidad de la propuesta y lo dota de flexibilidad para integrar datos de variada naturaleza en estos estudios.

3. Capítulo 3: Aplicación del método propuesto en la estratificación

En este capítulo se presentan los principales elementos relacionados con la validación del método de estratificación de territorios basado en Sistemas de Información Geográfica y medidas de similitud geométrica. También se discuten los resultados obtenidos a partir de la aplicación en casos de estudios, con el objetivo de comprobar que se obtienen grupos más compactos a través del método propuesto. Se utilizan índices de validación de grupos internos y externos para evaluar la competitividad de las soluciones. Se muestran los resultados de la dependencia espacial utilizando la I de Moran y por último se presentan los resultados de las pruebas estadísticas no paramétricas aplicadas.

3.1. Diseño de la validación

En la presente investigación la hipótesis plantea que: el desarrollo de un método de estratificación de territorios basado en SIG y medidas de similitud geométrica aplicado en la herramienta informática QGIS facilitará la obtención de grupos más compactos. Esta es del tipo hipótesis causal bivariada y se puede establecer la siguiente relación:

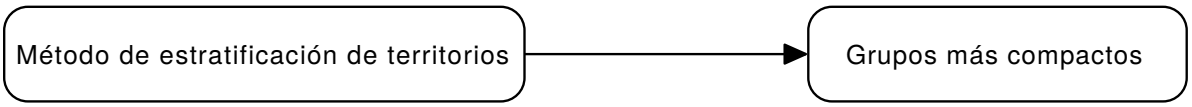


Figura 3.1: *Relación de la variable independiente y la dependiente, elaboración propia*

El diseño de la validación del método propuesto se basó en los siguientes elementos:

- El estudio de los métodos de validación de las tesis doctorales defendidas en los últimos cinco años en el tribunal de Computación.
- La literatura referida al tema, especialmente la relacionada con los métodos para validar investigaciones en el campo de la ingeniería de software y la minería de datos.

En la tabla 3.1 se muestra la medida de cada variable utilizada para corraborar la variable dependiente.

Tabla 3.1: *Operacionalización de la variable dependiente, elaboración propia*

Variable	Definición operacional
Más compacto	Índices de validación interna: miden la calidad de la solución en función de la distribución de las instancias por los agrupamientos, es decir, evalúan la separación que existe entre los clústeres y la compacidad que hay entre las instancias que pertenecen al mismo clúster. ▪ Silhouette: [-1,1]. Cuando tiende a –1 el agrupamiento no es bueno, 0 es indiferente y cuando tiende a 1 es bueno. ▪ Davies-Bouldin (DB) Un valor menor de DB indica una mejor solución de agrupamiento. Consecuentemente el número de grupos que minimiza el índice DB se toma como el óptimo. ▪ Calinski-Harabasz mide la separación en función de la distancia máxima entre los centros del clúster y la compacidad basada en la suma de las distancias entre los objetos y su centro del clúster. A mayor valor mejor solución.
	Índices de validación externa: Los índices externos de comparación son índices diseñados para medir la similitud entre dos particiones. Solo tienen en cuenta la distribución de los puntos en los diferentes grupos y no permiten medir la calidad de esta distribución, evalúan los agrupamientos en función de algún atributo externo. ▪ F-measure combina los conceptos de precisión y recuperación de la información. Los valores están dentro del intervalo [0-1] y los valores más grandes indican una mayor calidad de agrupamiento. ▪ Fowlkes-Mallows es la media geométrica de los coeficientes de precisión y recuperación. Un alto valor de este índice significa una mejor precisión. ▪ Jaccard los valores del índice Jaccard oscilan entre [0-1] y mayor valor indican la mejor validez del clúster, cero si no hay elementos que intercepten e igual a uno si todos los elementos interceptan. ▪ Rand index: Los valores del índice varían entre [0-1], mayor valor indica que todas las instancias de datos son agrupado correctamente y el clúster contiene sólo instancias puras.
	I de Moran se utiliza para la evaluación de la dependencia espacial a partir de la autocorrelación. Los resultados varían entre el –1 y 1, representando las correlaciones mínimas (máxima dispersión) y máximas (máxima concentración) respectivamente y el cero significa un patrón espacial totalmente aleatorio.

Para la validación se realizaron dos casos de estudios. El primero con el objetivo de verificar la competitividad del

método propuesto con otros trabajos reportados en la literatura. En el segundo caso de estudio se pretende evaluar la validez del método propuesto para abordar estudios estratificados desde la dinámica espacial en correspondencia con la primera ley de la geografía. En los siguientes epígrafes se describen los casos de estudio y se discuten los resultados obtenidos.

3.2. Estratificación de territorio basada en indicadores del año 2001

Para valorar los resultados de la solución propuesta se decide aplicar un caso de estudio, en correspondencia con el trabajo realizado por Companioni [2], en el cual la autora realiza una estratificación de territorio en Cuba para analizar el comportamiento de indicadores de salud. Utiliza la división política-administrativa del año 1976, en la cual existía una composición de 14 provincias, selecciona cuatro indicadores y un total de 26 variables del Anuario Estadístico del año 2001.

Los indicadores seleccionados fueron: mortalidad infantil por cada 1000 nacidos vivos, mortalidad infantil de los niños menores de cinco años por cada 1000 nacidos vivos, mortalidad por enfermedades del corazón por cada 100 000 habitantes, mortalidad por tumores malignos por cada 100 000 habitantes, mortalidad por enfermedades cerebrovasculares por cada 100 000 habitantes, mortalidad por influenza y neumonía por cada 100 000 habitantes, mortalidad por accidentes por cada 100 000 habitantes, mortalidad perinatal por cada 1000 nacidos vivos, mortalidad por enfermedades infecciosas y parasitarias por cada 100 000 habitantes, mortalidad materna por cada 100 000 nacidos vivos, incidencia de tuberculosis por cada 100 000 habitantes, incidencia de hepatitis por cada 100 000 habitantes, incidencia de diabetes por cada 100 000 habitantes. Incidencia de hipertensión por cada 100 000, incidencia de asma por cada 100 000 habitantes, incidencia de bajo peso al nacer, consultas médicas por habitante, ingresos por cada 100 habitantes, camas de asistencia por cada 1000 habitantes, consultas de puericultura por habitante, consultas de pediatría por habitante, densidad poblacional, población mayor de 60 años, población menor de un año, población menor de 15 años, natalidad por cada 1000 habitantes.

Para realizar el estudio descrito utilizando el método propuesto en la presente investigación, se emplea el algoritmo de agrupamiento K-Means, el número de estratos se fija en cuatro y se seleccionan los indicadores en correspondencia con el trabajo realizado por Companioni [2]. Luego se realiza una estratificación para cada función de distancia y de similitud geométrica declarada en esta investigación. Se realiza a continuación una evaluación del comportamiento del agrupamiento utilizando índices de validación de agrupamientos internos reportados en la literatura como Davies-Bouldi, Calinski-Harabasz y Silhouette [169, 194] para tener la cercanía de los elementos del clúster y la separabilidad que es el cálculo de la distancia entre dos grupos diferentes, obteniéndose que tan distintos son dos grupos.

A través de la interfaz que se muestra en la figura 2.10 se escogen las fuentes tanto temáticas como cartográficas.

Se escoge como base para la estratificación una capa de polígonos obtenida desde la IDERC [1] con la división política-administrativa vigente para el año 2001.

Una vez seleccionadas las fuentes de datos para el estudio, se procede a la configuración de los parámetros, la selección de los territorios y los indicadores que se utilizarán desde la ventana que se muestra en la figura 3.2. En este estudio se analizan todas las provincias y los indicadores, además se clasifican estos últimos en función de su aporte al riesgo a partir de la ecuaciones 2.8 y 2.7.

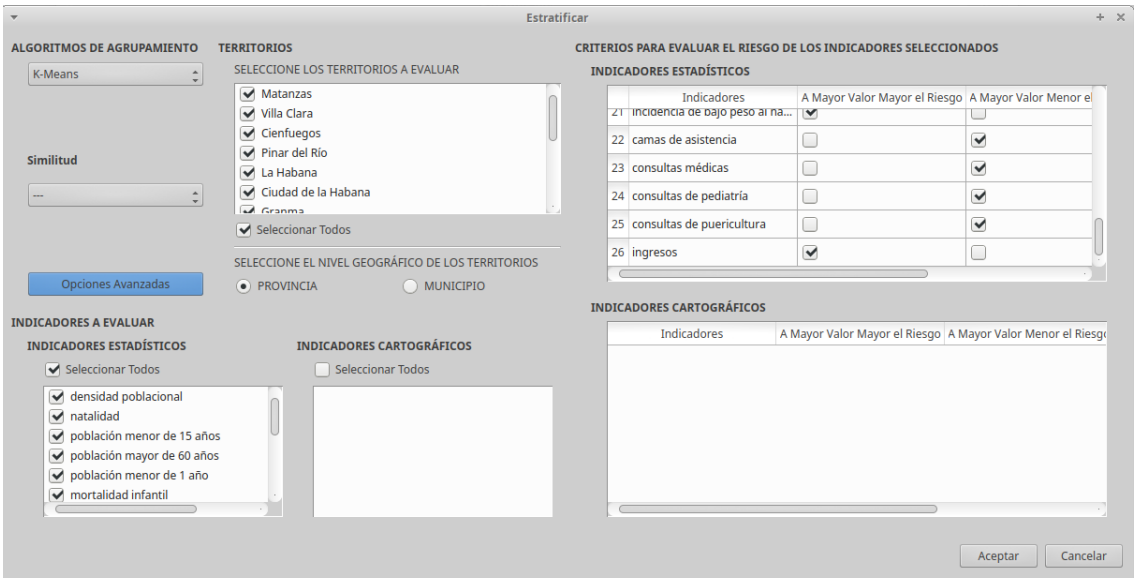


Figura 3.2: *Selección de indicadores y parámetros para la estratificación, tomada de XANGEO*

3.3. Resultados de la estratificación de territorio basada en indicadores del año 2001

A partir de los territorio y los indicadores seleccionados se obtienen los grupos y se visualizan en un mapa temático. En la figura 3.3 se muestra el resultado de la estratificación.

[1] Infraestructura de Datos Espaciales de la República de Cuba

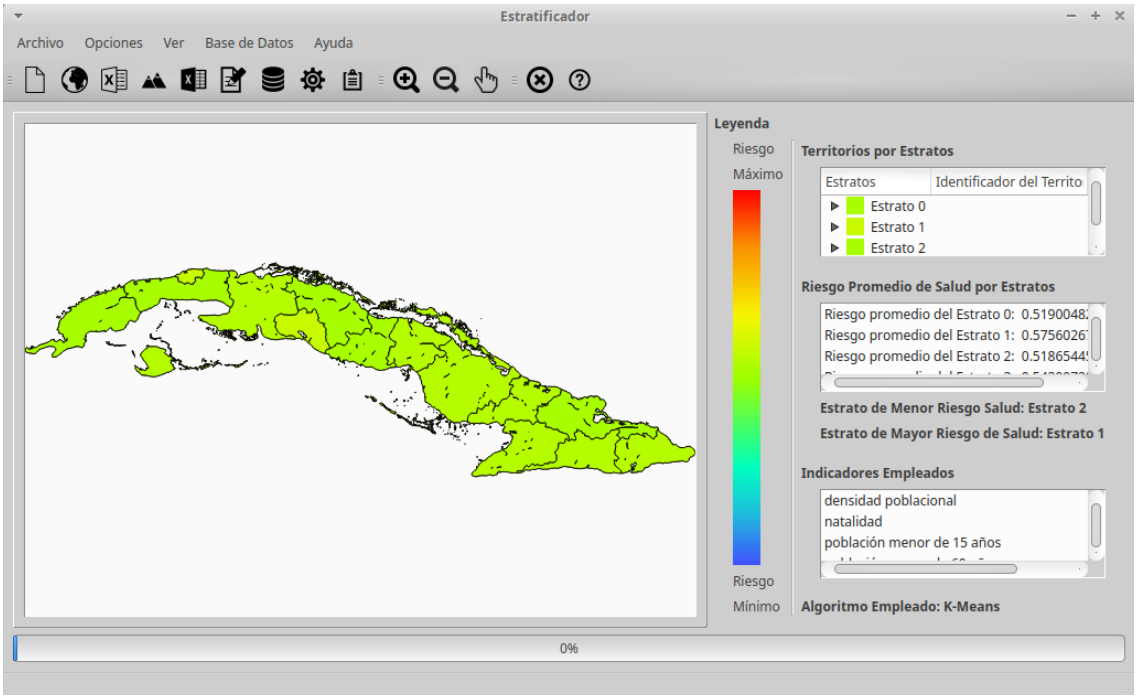


Figura 3.3: *Visualización de la estratificación, elaboración propia*

Luego de realizados los estudios estratificados utilizando cada una de las funciones de distancia y similitud propuestas (distancia, conectividad y tamaño), se procede a calcular los índices de validación interno para cada estudio. Se incluye también el cálculo para los estudios presentado por Companioni [2] y Pérez [3], este último forma parte de los resultados de esta investigación.

En la tabla 3.2 se muestran los valores obtenidos, se evidencia que los resultados utilizando el método propuesto en esta investigación son competitivos y que para el caso del Silhouette los resultados son superiores.

Tabla 3.2: *Validación con índices internos, elaboración propia*

Índice	Davies_Bouldin	Calinski_Harabaz	Silhouette_Score
Companioni 2005	1.991	2.182	-0.029
Pérez et al 2016	1.748	2.484	0.019
Conectividad	1.810	1.827	0.005
Distancia	1.903	1.199	0.012
Tamaño	1.800	1.247	0.013

Luego se procede a comparar los resultados obtenidos a partir de la utilización de las tres funciones propuestas para evaluar los índices externos precisión (P), *Jaccard* (J), *Fowlkes Mallows* (F&M), *Rand_Index* (Rand), *Bray*

Curtis (BC), *V_measure* (VM), *Mutual_Info* (MI), *Completeness* (C) y *Homogeneity* (H) teniendo como referencia los estudios Companioni [2] y Pérez [3].

En la tabla 3.3 se muestran los valores obtenidos teniendo como referencia al trabajo de [2] y se evidencia un mejor desempeño de la función conectividad en cuanto a precisión y al índice de *Jaccard*. Para el resto de índices los mejores resultados de forma general los obtiene la función basada en el tamaño como se puede ver en la figura 3.4.

Tabla 3.3: *Evaluación de índices a partir de Companioni [2], elaboración propia*

Comparaciones	P	J	F&M	Rand	BC	VM	MI	C	H
Conectividad	0.45	0.21	0.28	0.54	0.27	0.22	0.25	0.24	0.20
Distancia	0.10	0.07	0.21	0.58	0.48	0.22	0.27	0.22	0.22
Tamaño	0.33	0.14	0.31	0.60	0.36	0.37	0.44	0.38	0.36

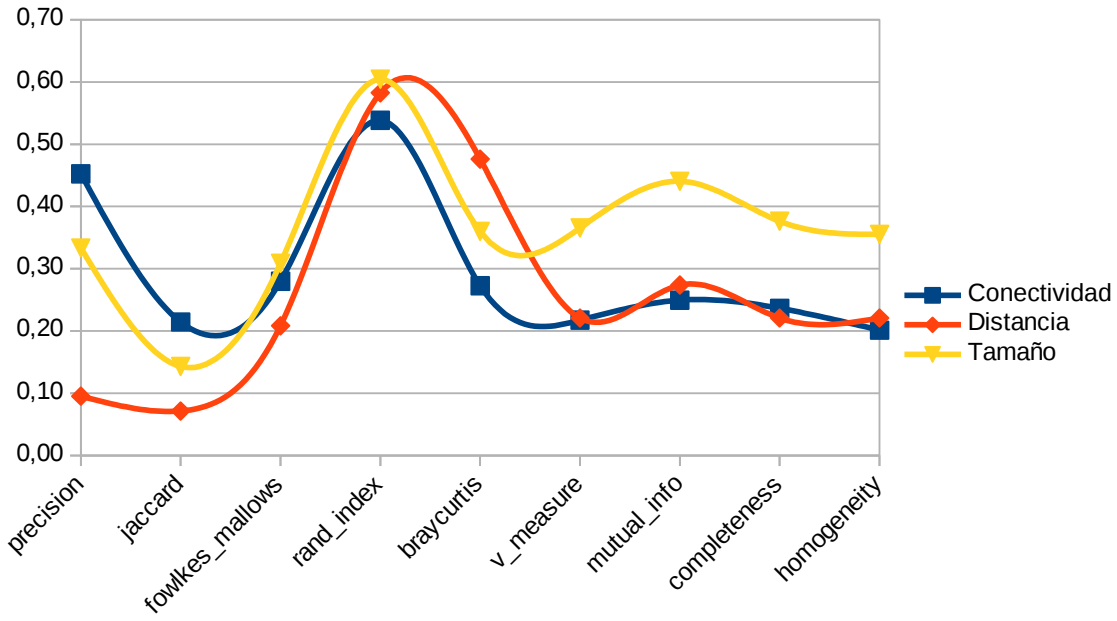


Figura 3.4: *Resultados de evaluación de las métricas con referencia a Companioni [2], elaboración propia*

La tabla 3.4 de forma general muestra mejores resultados para la comparación con Pérez [3] que con Companioni [2]. De igual forma se observan mejores resultados en cuanto a precisión y al índice de *Jaccard* para la conectividad y para el resto, el mejor desempeño lo obtiene la función basada en la distancia entre polígonos. Para todos los casos los mejores resultados se obtienen al evaluar el índice *Rand_Index*.

Tabla 3.4: *Evaluación de métricas a partir de Pérez[3], elaboración propia*

Comparaciones	P	J	F&M	Rand	BC	VM	MI	C	H
Conectividad	**0.50**	**0.29**	0.29	0.56	0.25	0.31	0.36	0.34	0.28
Distancia	0.14	0.14	0.30	**0.62**	0.40	0.40	0.50	0.41	0.40
Tamaño	0.38	0.21	0.28	0.60	0.29	0.31	0.37	0.29	0.32

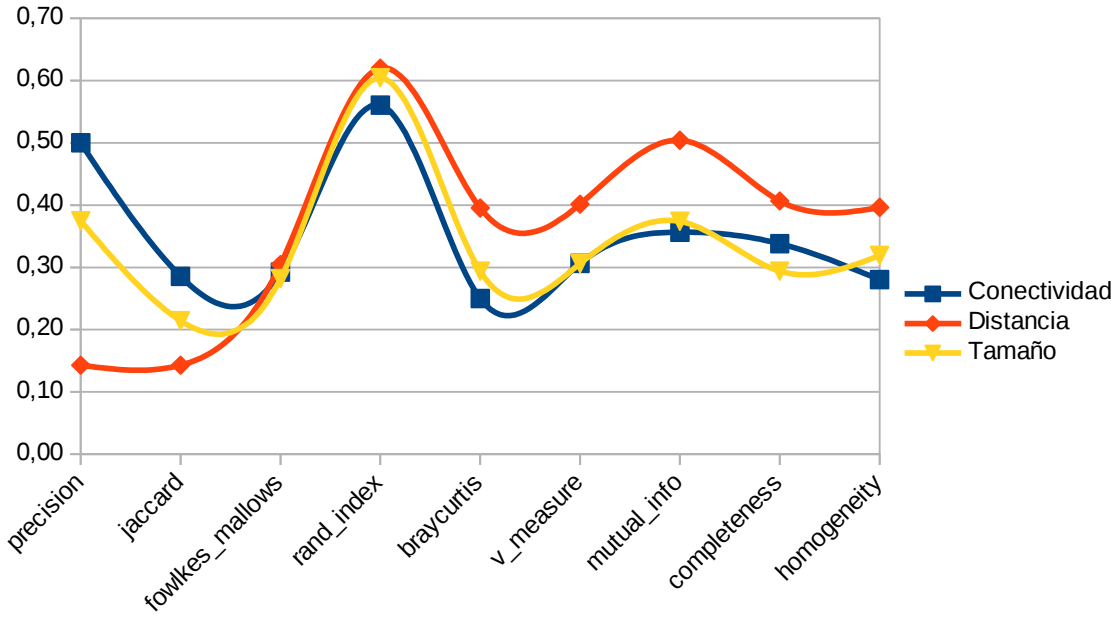


Figura 3.5: *Resultados de evaluación de las métricas con referencia a Pérez [3], elaboración propia*

En la figura 3.6 se observa que siempre los mejores valores se obtienen al tener como referencia el trabajo de Pérez [3], lo que está en correspondencia con los resultados obtenidos a partir de los criterios de evaluación internos.

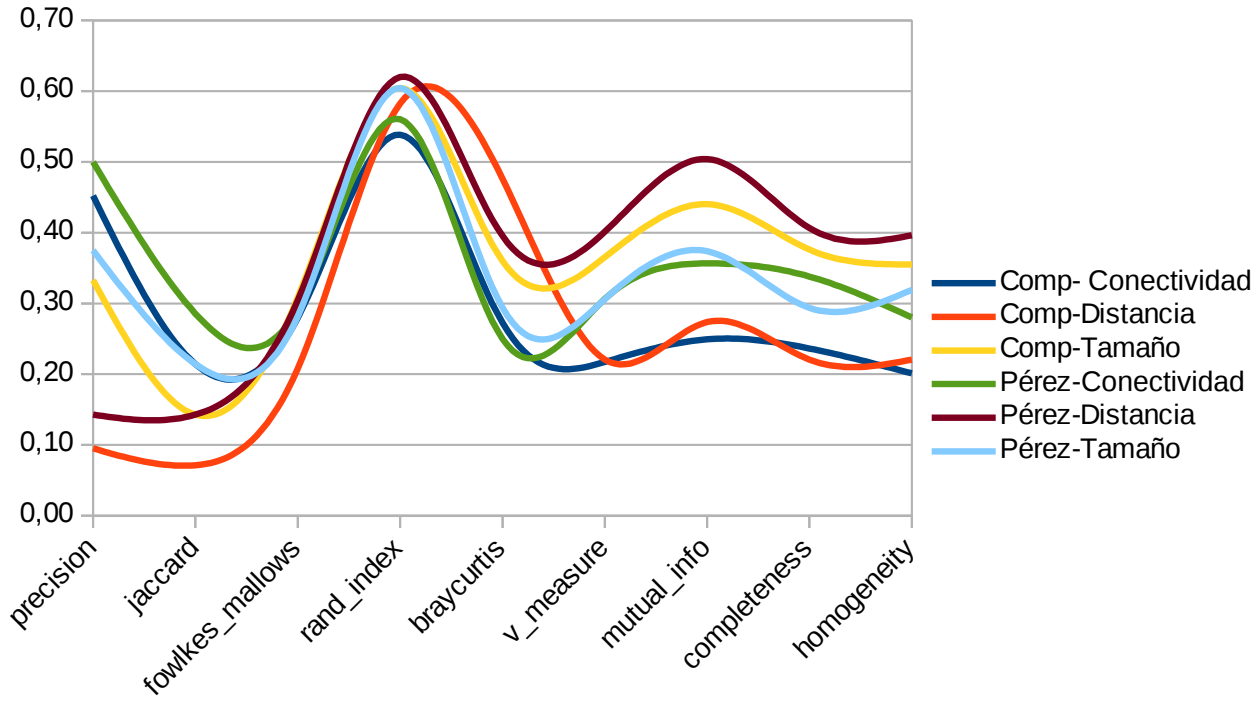


Figura 3.6: *Resultados de evaluación de las métricas con referencia a Companioni [2] y Pérez [3], elaboración propia*

3.4. Estratificación de territorios según las diez principales causas de muerte en el año 2016

Para valorar los resultados de la solución propuesta se decide aplicar otro caso de estudio, en correspondencia con el trabajo realizado por Pérez [20], en el cual se utiliza la división política-administrativa del año 2011, por ello seleccionan 15 provincias y el municipio especial Isla de la Juventud. Obteniéndose una capa vectorial desde la IDERC con los polígonos que representan a cada territorio escogido para el análisis. Parten de la hipótesis de relación de las enfermedades con el espacio y seleccionan como variables las diez principales causas de muerte de Cuba en el año 2016. Los indicadores de estas variables por territorios se obtuvieron del Anuario estadístico de salud del mencionado año.

Para realizar el estudio descrito utilizando el método propuesto en la presente investigación, se emplea el algoritmo de agrupamiento K-Means, el número de estratos se estima a partir de índices de validación de clúster, fundamentalmente el coeficiente de silueta. Luego se realiza una estratificación para cada función de distancia y de similitud geométrica declarada en esta investigación. Se realiza además una evaluación del comportamiento del agrupamiento utilizando índices de validación de agrupamientos internos reportados en la literatura como Davies-Bouldi, Calinski-Harabasz y Silhouette [169, 194] para tener la cercanía de los elementos del clúster y la separabilidad que es el calculo la distancia entre dos grupos diferentes, obteniéndose que tan distintos son dos

grupos.

Tabla 3.5: *Estimación de k a partir de índices internos, elaboración propia*

k	Davies_Bouldin	Calinski_Harabaz	Silhouette_Score
2	1.593	5.229	0.188
3	1.470	4.253	0.161
4	1.316	4.174	0.165
5	1.187	3.884	0.145
6	0.888	3.809	0.122
7	0.822	4.013	0.148
8	0.742	4.384	0.150

Al evaluar los resultados obtenidos y que se muestran en la tabla 3.5 se evidencia mejor desempeño para $k = 4$. Si se obtiene la curva de la función para el coeficiente de silueta se observa un punto de inflexión para $k = 4$, por lo que se toma cuatro como la cantidad de grupos a formar y está en correspondencia con Pérez [20].

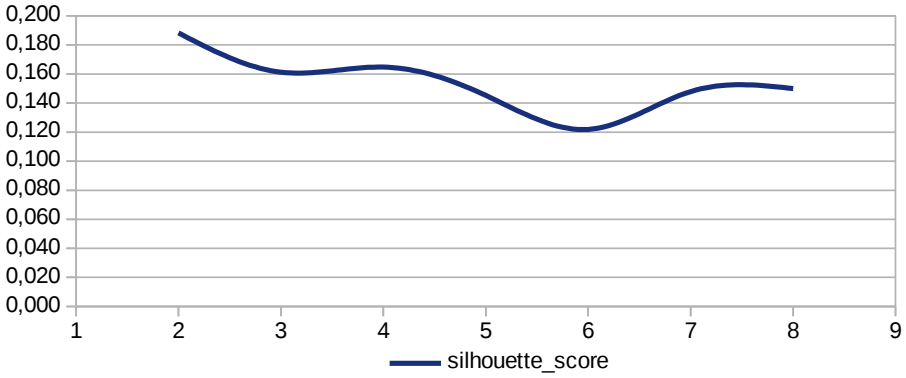


Figura 3.7: *Coeficiente de silueta, elaboración propia*

3.5. Resultados de la estratificación de territorios según las principales causas de muerte en el 2016

Los resultados de los índices de validación internos para los estudios realizados se muestran en la figura 3.8. Primeramente se observa que con las funciones de distancia geométricas se obtienen siempre grupos más compactos que solo con la utilización de la componente temática. Para el resto de los índices los resultados son competitivos en comparación con la función temática y destaca el desempeño de la conectividad y la distancia en el espacio.

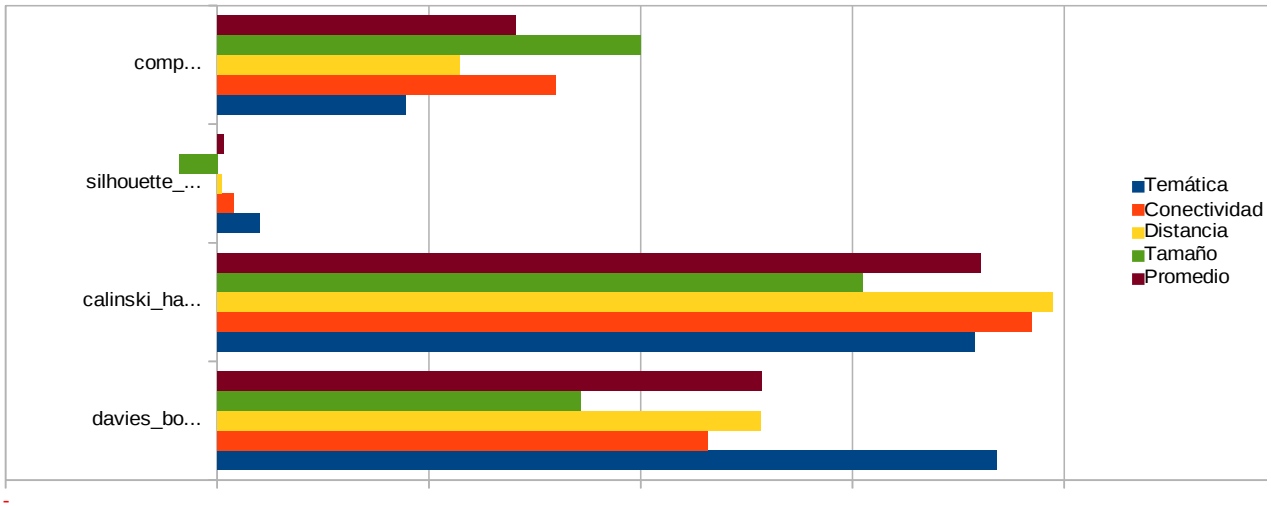


Figura 3.8: *Evaluación de índices de validación internos, elaboración propia*

Posteriormente se calculan los índices de validación tomando como referencia la temática, evidenciando un buen resultado en cuanto a la precisión por lo que las funciones de distancia geométricas para este estudio obtienen grupos más compactos sin afectar la precisión.

Tabla 3.6: *Resultados de las métricas de validación de clúster, elaboración propia*

Estratificación	P	J	F&M	Rand	V_M	MI	C	H
Conectividad	0.90	0.50	0.47	0.48	0.39	0.38	0.55	0.30
Distancia	0.79	0.50	0.40	0.54	0.31	0.33	0.36	0.27
Tamaño	0.90	0.44	0,.45	0.46	0.37	0.36	0.53	0.29

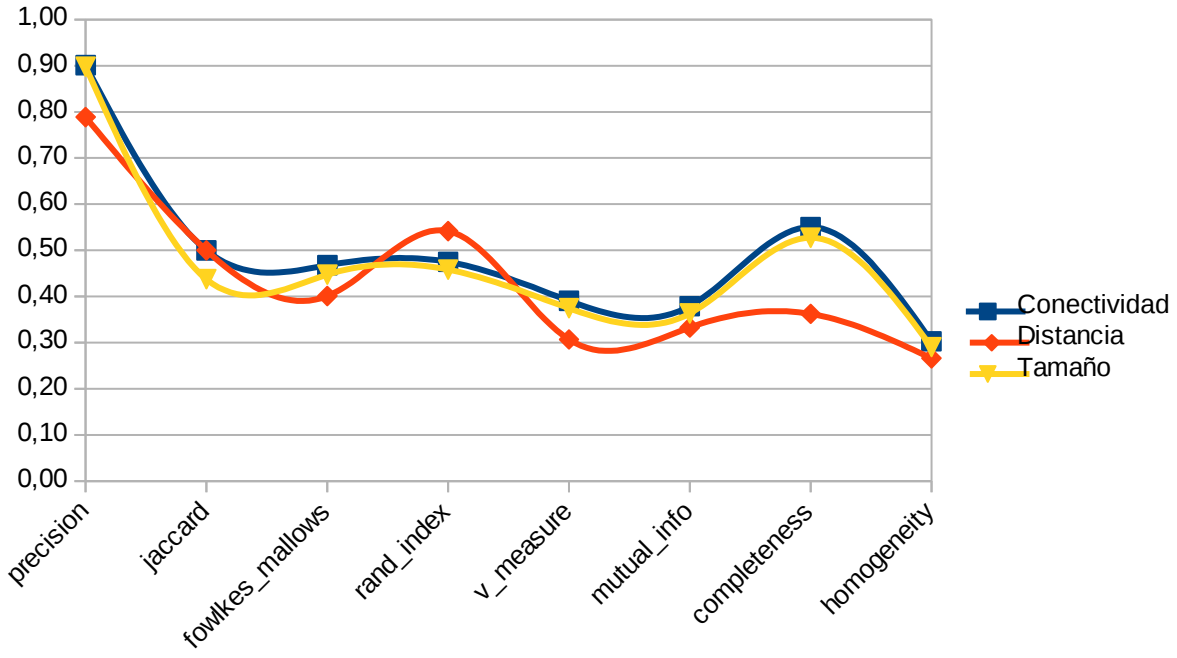


Figura 3.9: *Evaluación de métricas de validación de clúster, elaboración propia*

3.5.1. Análisis de resultados

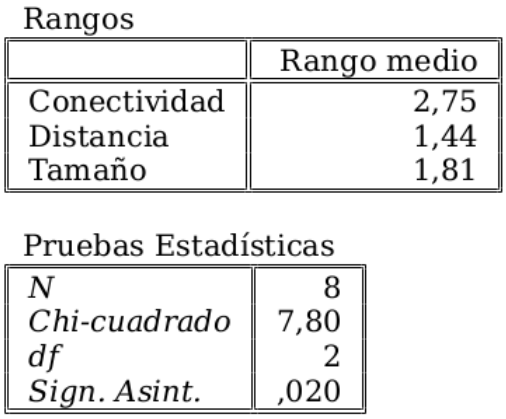
En la literatura consultada recomiendan emplear pruebas no paramétricas (no presuponen una distribución de probabilidad para los datos) basadas en rangos de Friedman seguida de comparaciones múltiples [195] para identificar pares de algoritmos que difieran significativamente. Se recomiendan cuando los tamaños de muestra son pequeños, especialmente cuando el número de instancias es menor que 30 [196]. En estos casos se emplea como parámetro de centralización la mediana, que es aquel punto para el que el valor de X está el 50 por ciento de las veces por debajo y el 50 por ciento por encima. Las ventajas del enfoque no paramétrico son: no promedia las medidas tomadas en diferentes conjuntos de datos, no asume normalidad de los medios muestrales y es robusto a los valores atípicos.

Friedman es una prueba de significación de hipótesis nula, por lo tanto, controla el error de Tipo I, es decir, la probabilidad de rechazar la hipótesis nula cuando es verdadera. Se le considera como un análisis de varianza no parámetrico para un diseño experimental en bloques. Por lo tanto hay que cumplir con dos suposiciones: se tiene k muestras relacionadas y la escala de medición de la variable a probar está al menos en escala ordinal [197].

Wilcoxon es considerada como una alternativa a la prueba de t (t de Student) para dos muestras pareadas. El procedimiento de ambas pruebas se basa en el cálculo de diferencias D_i=x_i– y_i entre pares de observaciones, pero en la prueba de Wilcoxon se asignan rangos a las diferencias. En esta prueba la hipótesis se plantea en torno a la mediana de la diferencias (Md), mientras que en la prueba de t se plantea sobre la media de diferencias (μ_D), en la

literatura consultada se enuncia una eficiencia de esta prueba del 95 por ciento para muestras pequeñas.

Se utilizan las pruebas no paramétricas porque a pesar de basarse en determinadas suposiciones, no parten de la base de que los datos analizados adoptan una distribución normal, el tamaño de la muestra es pequeño. De las pruebas disponibles en la literatura se aplica Friedman, que es una prueba para comprobar la igualdad de tratamientos en medidas repetidas, para contrastar la Hipótesis nula de igualdad entre esos tratamientos.



Rangos

	Rango medio
Conectividad	2,75
Distancia	1,44
Tamaño	1,81

Pruebas Estadísticas

N	8
Chi-cuadrado	7,80
df	2
Sign. Asint.	,020

Figura 3.10: *Resultados de la prueba estadística Friedman, elaboración propia*

El estadístico de prueba es igual a 7.80, por lo que se puede rechazar la hipótesis nula para niveles de significación superiores a 0.020. Al cinco por ciento de nivel de significación se rechaza la hipótesis de que no existen diferencias significativas y se concluye que al menos una de la medidas tiene un comportamiento diferente en la estratificación. Se realizan comparaciones dos a dos empleando la prueba de Wilcoxon de rangos con signo. Wilcoxon permite determinar si la diferencia entre la magnitud de las diferencias positivas entre los valores de las dos variables y la magnitud de las diferencias negativas es estadísticamente significativa. Los resultados obtenidos y que se muestran en la tabla 3.7 evidencian diferencias significativas solamente entra los criterios de conectividad y tamaño.

Tabla 3.7: *Resultados de la prueba de Wilcoxon, elaboración propia.*

	Conectividad – Tamaño	**Distancia – Tamaño**	**Conectividad – Distancia**
Z	-2.39	-1.05	-1.86
Sig. Asint.	0.017	0.293	0.063

3.5.2. Evaluación de la dependencia espacial del índice de riesgo

Para la evaluación de la dependencia espacial a partir de la autocorrelación se utiliza la I de Moran. La prueba de significación parte de las siguientes hipótesis:

- H_0 : no existe autocorrelación espacial. $H_0 : p_value > 0{,}05$
- H_a : existe autocorrelación espacial. $H_a : p_value < 0{,}05$

Al realizar el análisis exploratorio de los datos obtenidos en el caso de estudio la autocorrelación global del índice de riesgo obtiene un valor de 0.47. Como el resultado es positivo se puede establecer la hipótesis alternativa de existencia de dependencia espacial. La prueba de hipótesis arroja un p-value igual a 0.008, la autocorrelación es positiva y significativa por lo que existe dependencia espacial del riesgo obtenido.

Se procede luego a identificar si los grupos obtenidos por los diferentes clasificadores se encuentran formados en correspondencia con esta autocorrelación. El clasificador que utiliza la medida de similitud temática encuentra

Tabla 3.8: *Resultados de autocorrelación. Autocorrelación (valor P), elaboración propia*

Estrato	Temática	Conectividad	Distancia	Tamaño
1	0.86 (0.18)	0.63 (0.003)	0.52 (0.02)	0.64 (0.005)
2	un objeto	un objeto	un objeto	un objeto
3	0.5 (0.12)	un objeto	un objeto	un objeto
4	-0.08 (0.4)	-1.5 (0.001)	0.26 (0.32)	un objeto

dos estratos, el uno y el tres, que tienen autocorrelación positiva. Para estos se puede establecer la hipótesis de autocorrelación, sin embargo las pruebas para un nivel de certeza del cinco por ciento no permiten descartar la H_0 pues obtienen valores de 0.18 y 0.12 respectivamente. En el caso del estrato cuatro, se obtiene un índice negativo y por tanto no hay autocorrelación. El estrato dos está formado por un solo objeto por lo que no es posible aplicar la prueba.

Según el criterio de conectividad, el estrato uno obtiene un valor positivo y la prueba de hipótesis arroja un valor de 0.003 por lo que se puede rechazar la hipótesis nula para niveles de significación superiores a 0.003. Para el estrato cuatro se obtiene valor negativo por lo que se puede considerar la no existencia de autocorrelación, las pruebas de hipótesis también lo confirman. El resto de los estratos están conformados por un solo objeto.

Para el criterio de la distancia geométrica dos estratos obtienen valores de autocorrelación positivos, el uno y el cuatro, pero solo para el uno es significativo. Los estratos dos y tres están conformados por un solo objeto respectivamente. Finalmente para el criterio del tamaño, el estrato uno tiene una autocorrelación positiva y significativa y el resto de los estratos solo tiene un objeto.

Las evidencias reportadas en la tabla 3.8 permiten concluir que para este caso de estudio, donde en una primera instancia se pudo probar la existencia de dependencia espacial, los criterios propuestos identifican estratos con autocorrelación positiva.

3.6. Conclusiones del capítulo

En el capítulo se muestran los resultados obtenidos al aplicar el método propuesto en dos casos de estudios. Primeramente se realizó la estratificación de indicadores del año 2001 y se comparó con otros trabajos publicados y luego se desarrolló un estudio estratificado sobre las principales causas de muerte en Cuba durante el año 2016. La evaluación de estos casos de estudio y los resultados obtenidos permiten arrojar las siguientes conclusiones:

1. Para ambos estudios es posible incorporar medidas de distancia geométricas y mantener un desempeño competitivo en relación con otros trabajos publicados.
2. Al evaluar los índices de validación tanto internos como externos se pudo comprobar que si se incorporan medidas de distancia geométricas se obtienen grupos más compactos sin afectar la precisión de la clasificación.
3. El análisis exploratorio de datos espaciales arrojó evidencias que permiten considerar un mejor desempeño en clasificadores con los criterios propuestos para identificar estratos con dependencia espacial y por tanto más compactos.

CONCLUSIONES

Como resultado de la presente investigación se obtuvo un método para la estratificación de territorio basado en Sistemas de Información Geográfica y medidas de similitud geométrica, para obtener grupos más compactos. En función de los resultados obtenidos se arribó a las siguientes conclusiones:

1. A partir de la sistematización de los principales referentes teóricos que sustentan la presente investigación, se confirma que las propuestas para la estratificación reportadas en la literatura presentan limitaciones para la incorporación de la componente espacial en el proceso de estratificación de territorios.
2. La identificación de los constructos dentro de la minería de datos geoespaciales facilitó la descripción adecuada del problema y su solución a partir del método propuesto, que además integra los enfoques aportados en investigaciones precedentes con relación a estudios estratificados y la componente espacial de los datos en el espacio de solución del problema.
3. La selección de criterios de similitud entre polígonos en función de las características de estudios de espacialidad en salud, facilitó el diseño de medidas que integran la componente espacial y pueden describir relaciones espaciales sobre objetos.
4. La instanciación del método propuesto a partir del desarrollo de un componente para el SIG QGIS permitió la evaluación concreta del método a través de la realización de dos casos de estudios. Esta evaluación, a partir del análisis exploratorio de datos espaciales, arrojó evidencias que permiten considerar un mejor desempeño de la propuesta para identificar estratos con dependencia espacial y por tanto en correspondencia con la primera ley de la geografía.
5. Al evaluar los índices de validación tanto internos como externos se pudo comprobar que si se incorporan medidas de distancia geométricas se obtienen grupos más compactos sin afectar la precisión de la clasificación.

RECOMENDACIONES

1. Trabajar temas asociados al tratamiento de la incertidumbre, el ruido y ausencia de información en este proceso, además de la integración de otras técnicas de análisis y simulación espacial en una plataforma que impacte favorablemente en los estudios salubristas que se realizan en Cuba.
2. Abordar la percepción geosocial del riesgo a partir del análisis de datos georreferenciados en las redes sociales.

REFERENCIAS BIBLIOGRÁFICAS

[1] Pena Suarez, A. Modelo para la Caracterización del Delito en la Ciudad de Bogotá, Aplicando Técnicas de Minería de Datos Espaciales. *Repositorio institucional: UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS*, julio 2017. Disponible en: http://repository.udistrital.edu.co/bitstream/11349/6519/1/Pe%c3%b1aSuarezAlfonso2017.pdf.

[2] Companioni, Y. B. Construcción de tipologías: metodología de análisis para la estratificación según indicadores de salud. *Reporte Técnico de Vigilancia*, 2005. Disponible en: http://www.bvs.sld.cu/uats/rtv_files/2005/bombino.htm.

[3] Pérez Betancourt, Y.; Betancourt, Y. G. P.; Polanco, L. G.; Pérez, R. M. & Vega, Y. T. Estratificación de territorios basada en indicadores de salud sobre el Sistema de Información Geográfica QGIS. *Revista Cubana de Ciencias Informáticas*, 10(0):163–175, May 2016. Disponible en: http://rcci.uci.cu/?journal=rcci&page=article&op=view&path[]=1374.

[4] Cerda L, J. & Valdivia C, G. John Snow, la epidemia de Colera y el nacimiento de la epidemiología moderna. *Revista chilena de infectologia*, 24:334, agosto 2007. Disponible en: https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0716-10182007000400014&nrm=iso.

[5] Marco Antonio, S. R.; Elsa Raquel, H. L. & Rodrigo, S. H. ARTÍCULO HISTÓRICO. Dr. John Snow, Padre de la Epidemiología Moderna. *Revista Médica Carrionica*, 4(3), 2017. Disponible en: http://cuerpomedico.hdosdemayo.gob.pe/index.php/revistamedicacarrionica/article/view/184/0.

[6] Rodriguez Morales, A. J.; Cardenas Giraldo, E. V.; Arias, M.; Cindy P, Guerrero Matituy, E. A.; Bedoya Arias, Juan E.and Ramirez Jaramillo, V. & Villamil Gomez, W. E. Mapping chikungunya fever in municipalities of one coastal department of Colombia (Sucre) using geographic information system (GIS) during 2014 outbreak: Implications for travel advice. *Travel Medicine and Infectious Disease*, Volume 13(3):256–258, may 2015. doi:https://doi.org/10.1016/j.tmaid.2015.05.007.

[7] Patrick, S. W.; Davis, M. M.; Lehmann, C. U. & Cooper, W. O. Increasing incidence and geographic distribution of neonatal abstinence syndrome: United States 2009 to 2012. *Journal of Perinatology*, page 650–655, july 2015. Disponible en: https://www.nature.com/articles/jp201536, doi:https://doi.org/10.1038/jp.2015.36.

[8] Frenk, J. *La salud de la población.: Hacia una nueva salud pública*. Fondo de Cultura Económica, 2016.

[9] Metzger, M. J.; H., B. R. G.; G., J. R. H.; A., M. C. & W., W. J. A climatic stratification of the environment of Europe. *Global Ecology and Biogeography*, 14(6):549–563, 2005. Disponible en: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1466-822X.2005.00190.x, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1466-822X.2005.00190.x, doi:10.1111/j.1466-822X.2005.00190.x.

[10] Metzger, M. J.; H., B. R. G.; G., J. R. H.; Roger, S.; Antonio, T. & Robert, Z. A high-resolution bioclimate map of the world: a unifying framework for global biodiversity research and monitoring. *Global Ecology and Biogeography*, 22(5):630–638, December 2012. Disponible en: https://onlinelibrary.wiley.com/doi/abs/10.1111/geb.12022, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/geb.12022, doi:10.1111/geb.12022.

[11] González Polanco, L. & Pérez Betancourt, G. La minería de datos espaciales y su aplicación en los estudios de salud y epidemiología. *Revista Cubana de Información en Ciencias de la Salud*, 24(4):482–489, 2013. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2307-21132013000400010.

[12] Acosta, H. D.; Díaz, S. M.; Buergo, D. R.; Galindo, M. V. & Acosta, M. S. Estratificación del bajo peso al nacer desde un enfoque de determinantes sociales. *Revista Finlay*, 3(1), 2013.

[13] Delgado Acosta, H.; Gonzalez Moreno, L.; Valdes Gomez, M.; Hernandez Malpica, S.; Montenegro Calderon, T. & Rodriguez Buergo, D. Estratificación de riesgo de tuberculosis pulmonar en consejos populares del municipio Cienfuegos. *MediSur*, 13(2):275–284, 2015. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1727-897X2015000200005.

[14] Ariadna, C. M. & María del Carmen, P. B. Clasificación del Territorio Nacional Según un índice de Condiciones de Vida, Cuba 2014. In *Convención Salud 2015*, 2015. Disponible en: http://www.convencionsalud2015.sld.cu/index.php/convencionsalud/2015/paper/viewPaper/596.

[15] Remigio, S. P.; Yanelis, G. V.; Idania, R. M.; Damaris, G. T. & Eduardo, S. L. Riesgo de aparicion de eventos de enfermedad diarreica aguda en provincia Guantánamo. Estratificación epidemiológica. *Revista Información Científica*, 89(1):123–135, 2015. Disponible en: http://www.revinfcientifica.sld.cu/index.php/ric/article/view/271.

[16] Martin, A. C. Estratificación de territorios según condiciones de vida como expresión de las desigualdades sociales en salud. *Revista Cubana de Medicina General Integral*, 33(3), 2017. Disponible en: http://revmgi.sld.cu/index.php/mgi/article/view/385.

[17] Castillo Salgado, C. Epidemiological risk sratification of malaria in the Americas. *Memorias do Instituto do Instituto Oswaldo Cruz*, 87:120, 1992. Disponible en: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0074-02761992000700017&nrm=iso.

[18] Quesada Aguilera, J. A.; Quesada Aguilera, E. & Rodríguez Socarras, N. Diferentes enfoques para la estratificación epidemiológica del dengue. *Revista Archivo Médico de Camagüey*, 16(1):109–123, February 2012. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_abstract&pid=S1025-02552012000100014&lng=es&nrm=iso&tlng=es.

[19] Castillo Salgado, C. Geo-epidemiologic mapping in the new public health surveillance. The malaria case in Chiapas, Mexico, 2002. *Gaceta médica de Mexico*, 153(Supl. 2):S5—S12, 2017. Disponible en: https://doi.org/10.24875/GMM.M000001, doi:10.24875/gmm.m000001.

[20] Yadian, P. B.; Liset, G. P.; Juan, F. R. & Alcides, C. C. Propuestas para el análisis geoespacial en estudios salubristas. *Revista Cubana de Ciencias Informáticas*, 12(2), 2018. Disponible en: http://rcci.uci.cu/?journal=rcci&page=article&op=view&path%5B%5D=1731.

[21] Batista Moliner, R.; Coutin Marie, G.; Feal Cañizares, P.; González Cruz, R. & Rodríguez Milord, D. Determinación de estratos para priorizar intervenciones y evaluación en Salud Pública. *Revista Cubana de Higiene y Epidemiología*, 39(1):32–41, April 2001. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_abstract&pid=S1561-30032001000100005&lng=es&nrm=iso&tlng=es.

[22] Ngo, T. D.; Le, H. X.; Nguyen, H. M.; Nguyen, T. Q. N. Q.; Ho, T. D.; Nguyen, X. X.; Nguyen, A. Q.; Dinh, H. S.; Martin, N.; Ohrt, C. et al. MALARIA EPIDEMIOLOGICAL STRATIFICATION IN VIETNAM, 2014. In *AMERICAN JOURNAL OF TROPICAL MEDICINE AND HYGIENE*, page 276. AMER SOC TROP MED & HYGIENE 8000 WESTPARK DR, STE 130, MCLEAN, VA 22101 USA, 2015.

[23] Broderick, B.; Sears, K. & Stockley, D. International Leaders Influencing the Quality, Risk, and Safety Movement in Healthcare. *Influencing the Quality, Risk and Safety Movement in Healthcare: In Conversation with International Leaders*, page 1, 2015.

[24] Sultan, S.; Irfan, S. & Ashar, S. Acute Promyelocytic Leukemia: a Single Center Study from Southern Pakistan. *Asian Pacific Journal of Cancer Prevention*, 16(17):7893–7895, 12 2015.

[25] Kim, H.-S.; Chung, C.-K. & Kim, H.-K. Geo-spatial data integration for subsurface stratification of dam site with outlier analyses. *Environmental Earth Sciences*, 75(2):168, 2016.

[26] Resendes, A. P. d. C.; Silveira, N. A. P. R. d.; Sabroza, P. C. & Souza-Santos, R. Determination of priority areas for dengue control actions. *Revista de saude publica*, 44(2):274–282, 2010. doi:10.1590/S0034-89102010000200007.

[27] León Cabrera, P.; Fariñas Reinoso, A. T.; Galindo Reymond, K.; Prior García, A.; Dihigo Faz, T. & Núñez Valdés, L. Estratificación epidemiológica del riesgo de las enfermedades emergentes y reemergentes por Áreas de salud. Provincia de Matanzas. 2002-2006. *Revista Médica Electrónica*, 34:34 – 46, 02 2012. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1684-18242012000100004&nrm=iso.

[28] Colimon, K.-M. *Fundamentos de epidemiología.* 9789588843759. 3a ed. edition, 2018.

[29] Leo, Y.; Fleury, E.; Alvarez-Hamelin, J. I.; Sarraute, C. & Karsai, M. Socioeconomic correlations and stratification in social-communication networks. *Journal of The Royal Society Interface*, 13(125):20160598, 2016.

[30] Liu, Q.; Deng, M.; Shi, Y. & Wang, J. A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity. *Computers & Geosciences*, 46:296 – 309, 2012. Disponible en: http://www.sciencedirect.com/science/article/pii/S0098300411004419, doi:https://doi.org/10.1016/j.cageo.2011.12.017.

[31] Eduardo S. Martins, Marcos Ribeiro, J. L.-F. Clustering of spatial data for knowledge extraction. *IEEE.Information Systems and Technologies (CISTI), 2016 11th Iberian Conference on*, Information Systems and Technologies (CISTI), 2016 11th Iberian Conference on, july 2016. Disponible en: http://ieeexplore.ieee.org/abstract/document/7521626/, doi:978-1-5090-1226-8.

[32] Martin Ester, Hans-Peter Kriegel, J. S. Algorithms and Applications for Spatial Data Mining. *Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS, Taylor and Francis.*, 2001.

[33] Tanser, F. C. & Le Sueur, D. The application of geographical information systems to important public health problems in Africa. *International journal of health geographics*, 1(4):9, 2002. Disponible en: http://ij-healthgeographics.biomedcentral.com/articles/10.1186/1476-072X-1-4, doi:10.1186/1476-072X-1-4.

[34] Liu, Y. Geographical stratification and the role of the state in access to higher education in contemporary China. *International Journal of Educational Development*, 44:108–117, 2015.

[35] Fotheringham, S. & Rogerson, P. *Spatial analysis and GIS*. CRC Press, 2014.

[36] Cuellar Luna, L. & Gutietrrez Soto, T. Desarrollo de la geografía médica o de la salud en Cuba. *Revista Cubana de Higiene y Epidemiología*, 52:388 – 401, 12 2014. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1561-30032014000300011&nrm=iso.

[37] Orallo, J. H.; Quintana, M. J. R. & Ramírez, C. F. *Introducción a la Minería de Datos*. Pearson Prentice Hall, 2004.

[38] Idrees, A. M.; Ibrahim, M. H. & El Seddawy, A. I. Applying Spatial Intelligence for Decision Support Systems. *Future Computing and Informatics Journal*, IV, November 2018. Disponible en: http://www.sciencedirect.com/science/article/pii/S2314728818300977, doi:https://doi.org/10.1016/j.fcij.2018.11.001.

[39] Enrique, M. E.; Colas, L. A. B.; Vicente, P. D.; Liliam, C. L. & Doraida, R. S. Contaminación atmosférica y prevalencia de asma en Centro Habana. *Revista Cubana de Higiene y Epidemiologiá*, 39(1):5–15, ene.-abr 2001. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1561-30032001000100001&lng=es&nrm=iso.

[40] Luna, L. C. & Melián, M. G. El fluoruro en aguas de consumo y su asociación con variables geológicas y geográficas de Cuba. *Revista Panamericana de Salud Pública*, 14(5), mayo 2003.

[41] GALA GONZALEZ, A.; OROPESA GONZALEZ, L.; ARMAS PEREZ, L. & GONZALEZ OCHOA, E. Tuberculosis por municipios y sus prioridades: Cuba 1999-2002. *Revista Cubana de Medicina Tropical*, 58:0 – 0, 04 2006. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0375-07602006000100012&nrm=iso.

[42] Mercedes de los A, R. & Rodriguez, M. A. Los Sistemas de Información Geográfica: una herramienta para la estratificación en salud. *HYGEIA, Revista Brasileira de Geografia Médica e da Saúde*, 3(5), Dezembro 2007. Disponible en: http://www.seer.ufu.br/index.php/hygeia/article/view/16880.

[43] María, A. P. A. Utilización del sistema de información geográfica: para determinar el comportamiento territorial de los factores de riesgo que influyen en la morbilidad por hepatitis viral A en la cabecera municipal de Güines. *Tropical Geography Institute*, 2008. Disponible en: http://repositorio.geotech.cu/jspui/bitstream/1234/441/4/Utilizaci%c3%b3n%20del%20sistema%20de%20informaci%c3%b3n%20geogr%

c3%a1fica%20para%20determinar%20el%20comportamiento%20territorial%20de%20los%20factores%20de%20riesgo.pdf.

[44] Alegret Rodríguez, M.; Grau Abalo, R. & Rodríguez Rodríguez, M. El enfoque espacio-temporal-contextual en el estudio del VIH-SIDA. *Revista Cubana de Salud Pública*, 34:0 – 0, 09 2008. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-34662008000300004&nrm=iso.

[45] Sánchez Padrón, G.; Rodríguez, M.guez González, B.; García Pérez, M.; González Rodríguez, I. d. l. C. & Jiménez Prieto, Y. Dengue: estratificación espacial de riesgo. Área de salud xx aniversario. Municipio Santa Clara. Villa Clara. 2006-2007. *REDVET. Revista Electrónica de Veterinaria*, 11(03B), Marzo 2010. Disponible en: http://www.redalyc.org/html/636/63613140037/.

[46] Verdasquera Corcho, D.; Pérez, G.rez Soler, K.; Norales Mejía, A. D. & Váquez Pérez, A. Estratificación del riesgo de enfermar y morir por leptospirosis humana. *Revista Cubana de Medicina Tropical*, 65:191 – 201, 06 2013. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0375-07602013000200006&nrm=iso.

[47] Durá Noira, and Rodríguez, M. A.; Ramírez, E. B.; Cedrá, T. et al. Exploración espaciotemporal del riesgo de enfermar de leucemia aguda en niños. *Revista Cubana de Salud Pública*, 42(4), 2016.

[48] Olaya, V. Sistemas de información geográfica. *Recuperado de: http://www. icog. es/TyT/files/Libro_SIG. pdf*, 2014.

[49] Cliff, A. D. & Ord, K. Spatial Autocorrelation: A Review of Existing and New Measures with Applications. *Economic Geography*, 46(sup1):269–292, 1970. Disponible en: https://www.tandfonline.com/doi/abs/10.2307/143144, arXiv:https://www.tandfonline.com/doi/pdf/10.2307/143144, doi:10.2307/143144.

[50] Paelinck, J. H. P. & Klaassen, L. L. H. *Spatial econometrics*, volume 1. Saxon House, 1979.

[51] Anselin, L. & Kelejian, H. H. Testing for spatial error autocorrelation in the presence of endogenous regressors. *International Regional Science Review*, 20(1-2):153–182, 1997.

[52] Hay, S. I.; Battle, K. E.; Pigott, D. M.; Smith, D. L.; Moyes, C. L.; Bhatt, S.; Brownstein, J. S.; Collier, N.; Myers, M. F.; George, D. B. & Gething, P. W. Global mapping of infectious disease. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1614), 2013. Disponible en: http://rstb.royalsocietypublishing.org/content/368/1614/20120250, arXiv:http://rstb.royalsocietypublishing.org/content/368/1614/20120250.full.pdf, doi:10.1098/rstb.2012.0250.

[53] Betancourt, Y. G. P.; Polanco, L. G. & Rodríguez, J. P. F. Estratificación de territorios basada en indicadores de salud y medidas de similitud geométricas. *Ciencias técnicas. In: Ed. , Edacun-Redipe,* ., Diciembre 2017.

[54] BUSQUÉ, J. & MAESTRO, M. Y. J. S. Estratificación ambiental de Cantabria: metodología, resultados y aplicaciones de interés pascícola. *3ª Reunión Científica de la SEEP*, junio 2014. Disponible en: http://cifacantabria.org/Documentos/2014_Estratificacion%20ambiental%20de%20Cantabria_%20metodologia,%20resultados%20y%20aplicaciones%20de%20interes%20pascicola.pdf.

[55] Kaplan, R. M. & Litrownik, A. J. Some statistical methods for the assessment of multiple outcome criteria in behavioral research. *Behavior Therapy*, 8(3):383–392, 1977.

[56] Hair, J. F.; Anderson, R. E.; Tatham, R. L.; Black, W. C. et al. *Analisis multivariante*, volume 491. Prentice Hall Madrid, 1999.

[57] Peña, D. *Análisis de datos multivariantes*. McGraw-Hill España, 2013.

[58] Johnson, D. E. *Métodos multivariados aplicados al análisis de datos*. 2000, 2000.

[59] Pérez López, C. Técnicas de análisis multivariante de datos. *Aplicaciones con SPSS, Madrid, Universidad Complutense de Madrid*, 2004.

[60] Kinnear, T. C. & Taylor, J. W. *Investigación de mercados: un enfoque aplicado*. McGraw-Hill,, 1989.

[61] Korte, G. *The Gis Book*. OnWord Press, NY, USA, 15 edition, 2001. Disponible en: http://books.google.com.cu/books?id=_C6oPvJ5S_EC.

[62] Fu, P. & Sun, J. *Web GIS: principles and applications*. Esri Press, 2010.

[63] Malczewski, J. & Rinner, C. GIScience, Spatial Analysis, and Decision Support. In: Multicriteria Decision Analysis in Geographic Information Science. Advances in Geographic Information Science. *Springer, Berlin, Heidelberg*, February 2015. doi:https://doi.org/10.1007/978-3-540-74757-4_1.

[64] Faridi, M.; Verma, S. & Mukherjee, S. Integration of GIS, Spatial Data Mining, and Fuzzy Logic for Agricultural Intelligence. In *Soft Computing: Theories and Applications*, pages 171–183. Springer, 2018.

[65] Nicola, S. & Suzanne, M. Understanding the use of geographical information systems (GIS) in health informatics research: A review. *Journal of Innovation in Health Informatics*, 24(2):228 – 233, june 2018. Disponible en: https://hijournal.bcs.org/index.php/jhi/article/view/940, doi:http://dx.doi.org/10.14236/jhi.v24i2.940.

[66] Nakamura, T.; Nakamura, A.; Mukuda, K.; Harada, M. & Kotani, K. Potential accessibility scores for hospital care in a province of Japan: GIS-based ecological study of the two-step floating catchment area method and the number of neighborhood hospitals. *BMC health services research*, 17(1):438, 2017.

[67] Suprem, A.; Mahalik, N. & Kim, K. A review on application of technology systems, standards and interfaces for agriculture and food sector. *Computer Standards & Interfaces*, 35(4):355 – 364, 2013. Disponible en: http://www.sciencedirect.com/science/article/pii/S0920548912000955, doi:https://doi.org/10.1016/j.csi.2012.09.002.

[68] Wang, Y. Q. MeteoInfo: GIS software for meteorological data visualization and analysis. *Meteorological Applications*, 21(2):360–368, 2014. Disponible en: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/met.1345, arXiv:https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/met.1345, doi:10.1002/met.1345.

[69] Li, M.; Fang, L.; Huang, X. & Goh, C. A spatial–temporal analysis of hotels in urban tourism destination. *International Journal of Hospitality Management*, 45:34 – 43, 2015. Disponible en: http://www.sciencedirect.com/science/article/pii/S0278431914001753, doi:https://doi.org/10.1016/j.ijhm.2014.11.005.

[70] Tanser, F. C. & le Sueur, D. The application of geographical information systems to important public health problems in Africa. *International Journal of Health Geographics*, 1(1):4, Dec 2002. Disponible en: https://doi.org/10.1186/1476-072X-1-4, doi:10.1186/1476-072X-1-4.

[71] Shi, X. & Kwan, M.-P. Introduction: geospatial health research and GIS. *Annals of GIS*, 21(2):93–95, 2015. Disponible en: http://www.tandfonline.com/doi/full/10.1080/19475683.2015.1031204, doi:10.1080/19475683.2015.1031204.

[72] Pérez, S.; María, A.; Díaz Bernal, Z.; López Puig, P. & Gómez de Haz, H. Sensibilidad de género en el análisis de la situación de salud del modelo cubano de medicina familiar. *Revista Cubana de Salud Pública*, 41(2):268–289, 2015.

[73] Mas Bermejo, P. Desarrollo, tendencia actual y retos de la Epidemiología en Cuba. *Revista Cubana de Medicina Tropical*, 63:5 – 6, 04 2011. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0375-07602011000100001&nrm=iso.

[74] Hilda, D. A.; Sonia, M. D.; Delfin, R. B.; Galindo, M. V. & Acosta, M. S. Estratificación del bajo peso al nacer desde un enfoque de determinantes sociales. *Revista Finlay*, 3(1), 2013. Disponible en: http://www.revfinlay.sld.cu/index.php/finlay/article/view/171.

[75] Liu, Y. Geographical stratification and the role of the state in access to higher education in contemporary China. *International Journal of Educational Development*, 44:108–117, September 2015. ISSN 0738-0593. Disponible en: http://scielo.sld.cu/scielo.php?pid=S2227-18992018000200004&script=sci_arttext&tlng=en, doi:https://doi.org/10.1016/j.ijedudev.2015.08.003.

[76] Alegret Rodríguez, M.; Herrera, M. & Grau Abalo, R. Las técnicas de estadística espacial en la investigación salubrista: caso síndrome de Down. *Revista Cubana de Salud Pública*, 34(4):0–0, 2008. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-34662008000400003.

[77] Aronoff, S. *Geographical Information Systems: A management perspective*. WDL Publications, Otawa Canadá, 1989.

[78] Great Britain Committee of Enquiry into the Handling of Geographic Information and Great Britain Dept of the Environment. *Handling geographic information: report to the Secretary of State for the Environment of the Committee of Enquiry into the Handling of Geographic Information*. Number v. 1. H.M.S.O., 1987. Disponible en: http://books.google.com/books?id=Q2Eh8Rrfb7UC.

[79] Clarke, M. Geographical information systems and model based analysis: towards effective decision support systems. In *Proceedings of the GIS Summer Institute Kluwer*, Amsterdam, 1989.

[80] Cowen, D. Lectura en el Centro Nacional de Análisis e Información Geográfica. In *Lectura en el Centro Nacional de Análisis e Información Geográfica*, Universidad de California, 1989.

[81] Campbell, H. & Masser, I. *GIS In Organizations: How Effective Are GIS In Practice?* CRC Press, 1995.

[82] SWEENEY, M. W. Geographic information systems. Water Environment Research, 1999,. *Water Environment Research*, 71(5):551–556, 1999. Disponible en: https://www.ingentaconnect.com/content/wef/wer/1999/00000071/00000005/art00004., doi:10.2175/106143099X133631.

[83] Zias-Roe, S. *Equity and Inclusion in Planning: Engaging a Uniquely Abled Vulnerable Population in the Participatory Process*. PhD thesis, Prescott, May 2018.

[84] Delgado Fernandez, T. Infraestructura Cubana de Datos Geoespaciales: Una necesidad nacional para la integración y diseminación de datos geoespaciales. In *II Congreso Internacional Geomatica 2000*, La Habana, Febrero 2000.

[85] Pick, J. B. *Geographical Information Systems*, pages 1–4. American Cancer Society, 2015. Disponible en: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118785317.weom070053, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118785317.weom070053, doi:10.1002/9781118785317.weom070053.

[86] Yasobant, S.; Vora, K. S.; Hughes, C.; Upadhyay, A. & Mavalankar, D. V. Geovisualization: a newer GIS technology for implementation research in health. *Journal of Geographic Information System*, 7(01):20, February 2015. Disponible en: http://www.scirp.org/journal/jgis.

[87] Bivand, R. S.; Pebesma, E. & Gómez-Rubio, V. *Hello World: Introducing Spatial Data*, pages 1–16. Springer New York, New York, NY, 2013. Disponible en: https://doi.org/10.1007/978-1-4614-7618-4_1, doi:10.1007/978-1-4614-7618-4_1.

[88] Samad Rouhani, S. M. H. Creating a GIS based data bank of health facilities in Mazandarn province. *Life Science*, 10(9s):381-386] (ISSN:1097-8135).:54, 2013. Disponible en: http://www.lifesciencesite.com/lsj/life1009s/054_20440life1009s_381_386.pdf.

[89] Anguix, A. & Díaz, L. gvSIG: A GIS desktop solution for an open SDI. *Journal of Geography and Regional Planing*, 1(3):8, 2008.

[90] Rodríguez, R. V. USO DE SISTEMAS DE INFORMACIÓN GEOGRÁFICA LIBRES PARA LA PROTECCIÓN DEL MEDIO AMBIENTE. CASO DE ESTUDIO: MANIPULACIÓN DE MAPAS RÁSTER CON DATOS CLIMÁTICOS. *Universidad y Sociedad*, 10(2), febrero 2018. Disponible en: https://rus.ucf.edu.cu/index.php/rus/article/view/841/937.

[91] Blazek, R.; Neteler, M. & Micarelli, R. The new GRASS 5.1 vector architecture. In *Open source GIS - GRASS users conference*. University of Trento, Italy, 2002. Disponible en: http://www.ing.unitn.it/~grass/conferences/GRASS2002/proceedings/proceedings/pdfs/Blazek_Radim.pdf.

[92] Mark, R. T. D. & Marie, J. F. *Spatial Analys a guide for ecologists*. Cambridge University Press, 2014. Disponible en: https://books.google.com.cu/books?hl=es&lr=&id=s0lCBAAAQBAJ&oi=fnd&pg=PR11&dq=spatial+analysis+&ots=4_2NTcib5z&sig=HoPUV_zq_mp_tselAz5v8AlexEs&redir_esc=y#v=onepage&q=spatial%20analysis&f=false.

[93] Lior, R. et al. *Data mining with decision trees: theory and applications*, volume 81. World scientific, 2014.

[94] D A, A.; Z, W. & Y, Y. Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. *Applied Computing and Informatics*, 12(1):108, 2016. Disponible en: http://www.sciencedirect.com/science/article/pii/S221083271400026X, doi:https://doi.org/10.1016/j.aci.2014.10.001.

[95] Vaswani, K. & Karandikar, A. An Algorithm for Spatial Data Mining using Clustering. *International Journal of Computer & Mathematical Sciences*, 6, August 2017. Disponible en: http://academicscience.co.in/admin/resources/project/paper/f201708201503229995.pdf.

[96] Diana Elba, D. P. & Susana, G. a. R. Modelos geo-espaciales para la vigilancia local de la salud. *Revista Panamericana de Salud Pública*, 23(6):394–402., 2008. Disponible en: https://www.scielosp.org/pdf/rpsp/2008.v23n6/394-402/es.

[97] Pfeiffer, D.; Robinson, T. P.; Stevenson, M.; Stevens, K. B.; Rogers, D. J.; Clements, A. C. et al. *Spatial analysis in epidemiology*, volume 142. Oxford University Press New York, 2008.

[98] Pfeiffer, D. U. & Stevens, K. B. Spatial and temporal epidemiological analysis in the Big Data era. *Preventive veterinary medicine*, 122(1-2):213–220, 2015.

[99] Wu, X.; Zhu, X.; Wu, G.-Q. & Ding, W. Data mining with big data. *IEEE transactions on knowledge and data engineering*, 26(1):97–107, 2014.

[100] Raymond, T. N. & Jiawei, H. CLARANS: a method for clustering objects for spatial data mining. *IEEE Xplore*, 14(5), 2002.

[101] Han, J.; Fu, Y.; Wang, W.; Chiang, J.; Gong, W.; Koperski, K.; Li, D.; Lu, Y.; Rajan, A.; Stefanovic, N. et al. DBMiner: A System for Mining Knowledge in Large Relational Databases. In *KDD*, volume 96, pages 250–255, 1996.

[102] Wang, L. & Li, X. Spatial epidemiology of networked metapopulation: An overview. *Chinese Science Bulletin*, 59(28):3511–3522, 2014. Disponible en: http://link.springer.com/article/10.1007/s11434-014-0499-8.

[103] Li, D.; Wang, S. & Li, D. *Spatial Data Mining*. Springer, 2015.

[104] Koperski, K.; Adhikary, J. & Han, J. Spatial data mining: progress and challenges survey paper. In *Proc. ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, Canada*, pages 1–10. Citeseer, 1996.

[105] Koperski, K. & Han, J. Data mining methods for the analysis of large geographic databases. In *Proc. of 10th Annual Conf. on GIS, Vancouver, BC*. Citeseer, 1996.

[106] Perumal, M.; Velumani, B.; Sadhasivam, A. & Ramaswamy, K. Spatial data mining approaches for GIS–A brief review. In *Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2*, pages 579–592. Springer, 2015.

[107] Shishehgar, M.; Mirmohammadi, S. N. & Ghapanchi, A. R. A survey on data mining and knowledge discovery techniques for spatial data. *International Journal of Business Information Systems*, 19(2):265–276, 2015.

[108] Taha, A. Knowledge Discovery In GIS Data. *arXiv preprint arXiv:1601.07241*, 2016.

[109] Witten, I. H.; Frank, E.; Hall, M. A. & Pal, C. J. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[110] Li, D.; Wang, S.; Yuan, H. & Li, D. Software and applications of spatial data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 6(3):84–114, 2016.

[111] Ali, S. S.; Saeed, A.; Teh, Y. W. & Tutut, H. Big Data Clustering: A Review. *Springer, Cham*, 2014. Disponible en: https://link.springer.com/chapter/10.1007/978-3-319-09156-3_49, doi:https://doi.org/10.1007/978-3-319-09156-3_49.

[112] Jain, A.; Duin, R. & Mao, J. Statistical pattern recognition: a review. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4 – 37, Jan 2000. Disponible en: https://ieeexplore.ieee.org/abstract/document/824819/, doi:10.1109/34.824819.

[113] Salton, G. Automatic term class construction using relevance—A summary of work in automatic pseudoclassification. *Information Processing & Management*, 16(1):1 – 15, 1980. Disponible en: http://www.sciencedirect.com/science/article/pii/0306457380900023, doi:https://doi.org/10.1016/0306-4573(80)90002-3.

[114] DOUGLASS, R. A cluster-based approach to browsing large document collections. *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–329, 1992. Disponible en: https://ci.nii.ac.jp/naid/10011074138/en/.

[115] Dhillon, I. S.; Mallela, S. & Kumar, R. Enhanced Word Clustering for Hierarchical Text Classification. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 191–200, New York, NY, USA, 2002. ACM. Disponible en: http://doi.acm.org/10.1145/775047.775076, doi:10.1145/775047.775076.

[116] Zhai, C. & Lafferty, J. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, April 2004. Disponible en: http://doi.acm.org/10.1145/984321.984322, doi:10.1145/984321.984322.

[117] Xu, X.; Ester, M.; Kriegel, H.-P. & Sander, J. A distribution-based clustering algorithm for mining in large spatial databases. *Data Engineering, 1998. Proceedings., 14th International Conference on*, 23-27 Feb. 1998 1998. Disponible en: https://ieeexplore.ieee.org/abstract/document/655795/.

[118] Martin Ester, A.; Kriegel, F.-P. & Sander, J. Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support. *Data Mining and Knowledge Discovery*, 4(2-3):193–216, July 2000. Disponible en: https://link.springer.com/article/10.1023/A:1009843930701.

[119] Kolatch, E. et al. Clustering algorithms for spatial databases: A survey. *PDF is available on the Web*, pages 1–22, 2001.

[120] Fahad, A.; Alshatri, N.; Tari, Z.; Alamri, A.; Khalil, I.; Zomaya, A. Y.; Foufou, S. & Bouras, A. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing*, 2(3):267–279, 2014.

[121] Damaris Pascual, G. *Algoritmos de agrupamiento basados en densidad y variación de clusters*. PhD thesis, Universitat Jaume, marzo 2010. Disponible en: http://www.cerpamid.co.cu/sitio/files/DamarisTesis.pdf.

[122] T, S.; C, M. S. R. & K, V. N. A Survey on Clustering Techniques for Big Data Mining. *Indian Journal of Science and Technology*, 9(3), January 2016. Disponible en: http://www.indjst.org/index.php/indjst/article/viewFile/75971/66894, doi:10.17485/ijst/2016/v9i3/75971.

[123] Sajana, T.; Rani, C. S. & Narayana, K. A survey on clustering techniques for big data mining. *Indian Journal of Science and Technology*, 9(3), 2016.

[124] Fahim, A. Homogeneous Densities Clustering Algorithm. *Information Technology and Computer Science*, octuber 2018. Disponible en: ttp://www.mecs-press.org/, doi:10.5815/ijitcs.2018.10.01.

[125] Hartigan, J. A. Clustering algorithms. *Wiley*, 1975.

[126] Xu, R. & Wunsch, D. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

[127] Aggarwal, C. C. & Reddy, C. K. *Data clustering: algorithms and applications*. CRC press, 2013.

[128] Celebi, M. E. *Partitional clustering algorithms*. Springer, 2014.

[129] Popat, S. K. & Emmanuel, M. Review and comparative study of clustering techniques. *International journal of computer science and information technologies*, 5(1):805–812, 2014.

[130] Arora, P.; Varshney, S. et al. Analysis of k-means and k-medoids algorithm for big data. *Procedia Computer Science*, 78:507–512, 2016.

[131] Pérez-Ortega, J.; Almanza-Ortega, N. N.; Adams-López, J.; González-Gárcia, M.; Mexicano, A.; Saenz-Sánchez, S. & Rodriguez-Lelis, J. Improving the Efficiency of the K-medoids Clustering Algorithm by Getting Initial Medoids. In *World Conference on Information Systems and Technologies*, pages 125–132. Springer, 2017.

[132] Lopez Caviedes, M. A. Herramienta para la estratificación de municipios en zonas de riesgo para la salud. Master's thesis, UNIVERSIDAD AUToNOMA DEL ESTADO DE HIDALGO, Julio 2004. Disponible en: http://dgsa.uaeh.edu.mx:8080/bibliotecadigital/bitstream/handle/123456789/29/Herramienta%20para%20la%20estratificacion.pdf?sequence=1&isAllowed=y.

[133] Perreard, L.; Fan, C.; Quackenbush, J. F.; Mullins, M.; Gauthier, N. P.; Nelson, E.; Mone, M.; Hansen, H.; Buys, S. S.; Rasmussen, K. et al. Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay. *Breast Cancer Research*, 8(2):R23, 2006.

[134] Bock, H.-H. Clustering methods: a history of k-means algorithms. In *Selected contributions in data analysis and classification*, pages 161–172. Springer, 2007.

[135] Wan, X.; Wang, J.; Zhong, Y. & Du, Y. Dbh-clus: A hierarchal clustering method to identify pick-up/drop-off hotspots. In *International Conference on Intelligent Computing*, pages 330–341. Springer, 2015.

[136] Kumar, D.; Wu, H.; Lu, Y.; Krishnaswamy, S. & Palaniswami, M. Understanding urban mobility via taxi trip clustering. In *Mobile Data Management (MDM), 2016 17th IEEE International Conference on*, volume 1, pages 318–324. IEEE, 2016.

[137] Kumar, K. M. & Reddy, A. R. M. A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method. *Pattern Recognition*, 58:39–48, October 2016. Disponible en: https://www.sciencedirect.com/science/article/pii/S0031320316001035, doi:https://doi.org/10.1016/j.patcog.2016.03.008.

[138] Schubert, E.; Sander, J.; Ester, M.; Kriegel, H. P. & Xu, X. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, 42(3):19, 2017. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0169743912002249.

[139] Dudik, J. M.; Kurosu, A.; Coyle, J. L. & Sejdic, E. A comparative analysis of DBSCAN, K-means, and quadratic variation algorithms for automatic identification of swallows from swallowing accelerometry signals. *Computers in biology and medicine*, 59:10–18, April 2015. Disponible en: https://www.sciencedirect.com/science/article/pii/S0010482515000244#!, doi:https://doi.org/10.1016/j.compbiomed.2015.01.007.

[140] Benítez, I. Técnicas de agrupamiento para el análisis de datos cuantitativos y cualitativos. *España: Universidad Politécnica de Valencia, Departamento de Ingeniería de Sistemas y Automática*, 2005.

[141] Schwering, A. Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey. *Transactions in GIS*, 12(1):5–29, 2008.

[142] García, N. M. *Uso de la similitud semántica para la recuperación de información geoespacial*. PhD thesis, Universitat dÁlacant-Universidad de Alicante, 2015.

[143] Bonillo, M. L. *Razonamiento basado en casos aplicado a problemas de clasificación*. PhD thesis, Tesis de grado. Universidad de Granada, España, 2003.

[144] Fleitas, N. S.; Alvarado, T. R.; Lorenzo, M. M. G. & Riverol, A. MEDIDAS DE SIMILITUD PARA LOS COMPONENTES DE LA ONTOLOGIA DE LA UNIÓN ELECTRICA. *Informática 2016: III Conferencia Internacional en Ciencias Computacionales e Informáticas*, Marzo 2016. Disponible en: https://www.researchgate.net/publication/323583817.

[145] Fleitas, N. S.; Rdoríguez, R. C.; Lorenzo, M. M. G. & Quesada, A. R. Modelo de manejo de datos, con el uso de inteligencia artificial, para un sistema de información geográfica en el sector energético. *Enfoque UTE*, 7(3):95–109, 2016.

[146] Tversky, A. Features of similarity. *Psychological review*, 84(4):327, 1977.

[147] Thomas, M. S. & Mareschal, D. Connectionism and psychological notions of similarity. *American Journal of sociology*, 1997.

[148] Markman, A. B. & Gentner, D. Thinking. *Annual review of psychology*, 52(1):223–247, 2001.

[149] Nedas, K. A. & Egenhofer, M. J. Spatial similarity queries with logical operators. In *International Symposium on Spatial and Temporal Databases*, pages 430–448. Springer, 2003.

[150] Machado-García, N.; González-Ruiz, L. & Balmaseda Espinosa, C. Recuperación de objetos geoespaciales utilizando medidas de similitud semántica. *Revista Cubana de Ciencias Informáticas*, 8(2):132–144, 2014.

[151] Bedoya, J. *Aplicacion de distancias entre terminos para datos planos y jerarquicos*. PhD thesis, Tesis de Maestria. Universidad Politecnica de Valencia. España, Valencia, septiembre 2011. Disponible en: https://riunet.upv.es/bitstream/handle/10251/15874/01TFM.pdf?sequence=1.

[152] Edna, H. V. & Eléctrica, I. Algoritmo de clustering basado en entropía para descubrir grupos en atributos de tipo mixto, 2015.

[153] Rodríguez, J. E. R.; Blanco, E. A. R. & Camacho, R. O. F. Clasificación de datos usando el método k-nn. *revista Vínculos*, 4(1):4–18, 2007.

[154] Michel, M. D. & Deza, E. *Encyclopedia of Distances*. Springer, Springer Heidelberg New York Dordrecht London, third edition, 2014. Disponible en: https://link.springer.com/book/10.1007/978-3-662-44342-2, doi:10.1007/978-3-662-44342-2.

[155] Fonseca, F. T. Ontology-driven geographic information systems. *Wiley Online Library*, 2001.

[156] Montes, E. C. & García, M. d. C. M. Introducción a los Sistemas de Información Geográfica. In *Manual de Tecnologías de la Información Geográfica aplicadas a la Arqueología*, pages 21–78. Museo Arqueológico Regional, 2016.

[157] Manandhar, S. K. Efficient algorithms for clustering polygonal obstacles. *UNLV Theses, Dissertations, Professional Papers, and Capstones*, May 2016. Disponible en: https://digitalscholarship.unlv.edu/thesesdissertations/2704.

[158] Wang, J. F.; Zhang, T. L. & Fu, B. J. A measure of spatial stratified heterogeneity. *Ecological Indicators*, 67:256, 2016.

[159] Ramos, M. & Rivière, J. Presentación: estratificación social y nuevas desigualdades. *Encrucijadas-Revista Crítica de Ciencias Sociales*, 14:1401, 2017.

[160] Sánchez, F. J. Z.; Martínez, W. C.; Ovalle, R. I. A.; Barrera, M. A. A.; González, G. B. & Esparza, L. J. R. Una estratificación socioeconómica para comparar dos momentos del desarrollo en México: 1930-2010. *Economía Sociedad y Territorio*, I(56):259–289, 2018.

[161] Vinnakota, S. & Lam, N. S. Socioeconomic inequality of cancer mortality in the United States: a spatial data mining approach. *International journal of health geographics*, 5(1):9, 2006.

[162] Zhao, F.; Zhu, R.; Zhang, L.; Zhang, Z.; Li, Y.; He, M.; Zhou, Y.; Guo, J.; Zhao, G. & Jiang, Q. Application of satscan in detection of schistosomiasis clusters in marshland and lake areas. *Zhongguo xue xi chong bing fang zhi za zhi= Chinese journal of schistosomiasis control*, 23(1):28–31, 2011.

[163] Coleman, M.; Coleman, M.; Mabuza, A. M.; Kok, G.; Coetzee, M. & Durrheim, D. N. Using the SaTScan method to detect local malaria clusters for guiding malaria control programmes. *Malaria Journal*, 8(1):68, 2009.

[164] INEGI. *Estratificador INEGI. Manual de usuario*. INEGI, Aguascalientes, 1 edition, 2018. Disponible en: http://www3.inegi.org.mx/estratificador/assets/pdf/Guia_EstratificadorV1_1.pdf.

[165] Rodriguez, M. d. l. A. & Rodríguez, M. A. Los Sistemas de Información Geográfica: una herramienta para la estratificación en salud. *Hygeia*, 3(5), 2007.

[166] Alegret Rodríguez, M.; Grau Abalo, R. & Rodríguez Rodríguez, M. El enfoque espacio-temporal-contextual en el estudio del VIH-SIDA. *Revista Cubana de Salud Pública*, 34(3):0–0, 2008.

[167] Norma E, B. H.; Milagros, A. R. & Oscar, A. F. Análisis espacial de la morbimortalidad del cáncer de mama y cérvix. Villa Clara. Cuba. 2004-2009. *Revista Española de Salud Pública*, 87(1), febrero 2013. Disponible en: http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1135-57272013000100006.

[168] Durán Morera, N.; Alegret Rodríguez, M.; Batista Hernández, N.; Botello Ramírez, E.; Cedré Hernández, T. & Hernández González, G. Exploración espaciotemporal del riesgo de enfermar de leucemia aguda en niños. *Revista Cubana de Salud Pública*, 42:536–546, 2016. Disponible en: https://www.scielosp.org/scielo.php?pid=S0864-34662016000400536&script=sci_arttext&tlng=pt.

[169] Rendon, E.; Abundez, I.; Arizmendi, A. & Quiroz, E. M. Internal versus external cluster validation indexes. *International Journal of computers and communications*, 5(1):27–34, 2011.

[170] Desgraupes, B. Clustering indices. *University of Paris Ouest-Lab Modal'X*, 1:34, 2013.

[171] Zhang, S.; Yang, Z.; Xing, X.; Gao, Y.; Xie, D. & Wong, H. S. Generalized Pair Counting Similarity Measures for Clustering and Cluster Ensembles. *IEEE Access*, 5:16904–16918, 2017.

[172] Jose Maria, L. R.; Jorge, G. G.; Maria, M. B. & Santos, J. R. Aproximacion al indice externo de validacion de clustering basado en chi cuadrado. *ResearchGate*, Conference: CAEPIA 2018, At Granada, Spain, Octuber 2018. Disponible en: https://www.researchgate.net/publication/328578140.

[173] Jauhiainen, S. & Karkkainen, T. A Simple Cluster Validation Index with Maximal Coverage. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. ESANN, 2017.

[174] Christian, W.; Jan, B. & Richard, R. Comparing the performance of biomedical clustering methods. *Nature Methods*, September 2015. Disponible en: https://www.nature.com/articles/nmeth.3583#article-info, doi:https://doi.org/10.1038/nmeth.3583.

[175] Bichler, M. Design science in information systems research. *Wirtschaftsinformatik*, 48(2):133–135, 2006.

[176] Dresch, A.; Lacerda, D. P. & Antunes Jr, J. A. V. *Design science research: A method for science and technology advancement*. Springer, 2014.

[177] Venable, J.; Pries-Heje, J. & Baskerville, R. FEDS: a framework for evaluation in design science research. *European Journal of Information Systems*, 25(1):77–89, 2016.

[178] Peffers, K.; Tuunanen, T.; Gengler, C. E.; Rossi, M.; Hui, W.; Virtanen, V. & Bragge, J. The design science research process: a model for producing and presenting information systems research. In *Proceedings of the first international conference on design science research in information systems and technology (DESRIST 2006)*, pages 83–106. sn, 2006.

[179] Chen, Y.; Miao, D. & Wang, R. A rough set approach to feature selection based on ant colony optimization. *Pattern Recognition Letters*, 31(3):226–233, 2010.

[180] Shelke, G.; Sonawane, A.; Bongale, N.; Dhanwe, S. & Bhosale, S. A Feature Subset Selection Algorithm for High Dimensional Data Based on Fast Clustering. *Imperial Journal of Interdisciplinary Research*, 3(3), 2017. Disponible en: http://www.imperialjournals.com/index.php/IJIR/article/view/4223.

[181] Rejer, I. Genetic Algorithms for Feature Selection for Brain–Computer Interface. *International Journal of Pattern Recognition and Artificial Intelligence*, 29(05):1559008, April 2015. Disponible en: http://www.worldscientific.com/doi/abs/10.1142/S0218001415590089, doi:10.1142/S0218001415590089.

[182] Arafat, H.; Elawady, R. M.; Barakat, S. & Elrashidy, N. M. Using rough set and ant colony optimization in feature selection. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, 2(1), 2013. 00007. Disponible en: http://ijettcs.org/Volume2Issue1/IJETTCS-2013-02-25-077.pdf.

[183] Jensen, R. & Shen, Q. Fuzzy-rough data reduction with ant colony optimization. *Fuzzy sets and systems*, 149(1):5–20, 2005.

[184] Bello, R.; Nowe, A.; Gomezd, Y. & Caballero, Y. Using ACO and rough set theory to feature selection. *WSEAS Transactions on Information Science and Applications*, 2(5):512–517, 2005. 00023. Disponible en: https://www.researchgate.net/profile/Ann_Nowe2/publication/234810626_Using_ACO_and_rough_set_theory_to_feature_selection/links/0deec5346914cd1dfc000000.pdf.

[185] Tabakhi, S.; Moradi, P. & Akhlaghian, F. An unsupervised feature selection algorithm based on ant colony optimization. *Engineering Applications of Artificial Intelligence*, 32:112–123, 2014.

[186] Crockford, A.; Jamal-Hanjani, M.; Hicks, J. & Swanton, C. Implications of intratumour heterogeneity for treatment stratification. *The Journal of pathology*, 232(2):264–273, 2014. Disponible en: https://onlinelibrary.wiley.com/doi/abs/10.1002/path.4270.

[187] Eichstaedt, J. C.; Schwartz, H. A.; Kern, M. L.; Park, G.; Labarthe, D. R.; Merchant, R. M.; Jha, S.; Agrawal, M.; Dziurzynski, L. A.; Sap, M. et al. Psychological language on Twitter predicts county-level heart disease mortality. *Psychological science*, 26(2):159–169, 2015.

[188] Simental Ávila, J. & Pompa Garciía, M. Incendios forestales: autocorrelación espacial de topografía y temporalidad. *Ciencia UANL*, 19(77):41–45, 2016.

[189] Arce, X. C. M.; Lestegás, F. R. & Quintá, F. X. A. La cartografía temática como recurso didáctico en los procesos de enseñanza y aprendizaje de las ciencias sociales para educación primaria. *Revista Brasileira de Educação em Geografia*, 6(11):428–438, 2016.

[190] Fernandes, M. G. Ilustração e cartografia nos manuais escolares de Geografia, do ensino básico e secundário, em Portugal (séculos XIX e XX). *Universidad de Alicante*, 2016.

[191] Calle, T. & Luján-Mora, S. Importancia de Accesibilidad Web en Mapas Geográficos para la Educación. *Latin American Journal of Computing Faculty of Systems Engineering Escuela Politécnica Nacional Quito-Ecuador*, 2(3), 2015.

[192] Sommerville, I. *Software Engineering, 10th Edition*. Pearson Copyright © 2016, Mar 24, 2015.

[193] Pressman, R. S. *Ingeniería del software UN ENFOQUE PRÁCTICO. Octava EDICIÓN*. The McGraw-Hill, 2013.

[194] Hamalainen, J.; Jauhiainen, S. & Karkkainen, T. Comparison of Internal Clustering Validation Indices for Prototype-Based Clustering. *Algorithms*, 10(3):105, 2017.

[195] Benavoli, A.; Corani, G.; Mangili, F. & Zaffalon, M. A Bayesian nonparametric procedure for comparing algorithms. In *International Conference on Machine Learning*, pages 1264–1272, 2015. Disponible en: http://proceedings.mlr.press/v37/benavoli15.pdf.

[196] Bogdan, T.; Magdalena, S.; Zbigniew, T. & Tadeusz, L. Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms. *International Journal of Applied Mathematics and Computer Science*, 22(4):867 – 881, December 2012. Disponible en: https://content.sciendo.com/view/journals/amcs/22/4/article-p867.xml, doi:DOI:https://doi.org/10.2478/v10006-012-0064-z.

[197] Badii, M.; Guillen, A.; Araiza, L.; Cerna, E.; Valenzuela, J. & Landeros, J. Métodos no paramétricos de uso común. *Revista Daena (International Journal of Good Conscience)*, 7(1), 2012.