Universidad Autónoma de Coahuila

Centro de Investigaciones

Socioeconómicas

Tesis

Regionalización eficiente de mínima divergencia en información y aplicaciones

que presenta

José Refugio Reyes Valdés

como requisito parcial para obtener el grado de

Doctor en Economía Regional

Director de tesis:

Dr. Luis Gutiérrez Flores

Saltillo, Coahuila — Junio de 2016

Aquí, como puedes ver, para quedarte donde estás tienes que correr lo más rápido que puedas... Y si quieres ir a otro sitio, deberás correr, por lo menos, dos veces más rápido.

Alicia a través del espejo
Lewis Carroll

Dedicatoria

A José R. Reyes R. (Don Cuco)
y Héctor Reyes Valdés (La Bola)
... dondequiera que estén

A Magdalena Valdés Vela.

Agradecimientos

- A Magdalena Reyes (*Acúi*) siempre al pie del cañón peleando no solo las batallas propias sino las ajenas que finalmente las hace suyas también.
- A Magdalena Campos, siempre dinámica y alerta en el día a día.
- A Humberto Reyes por los consejos y jornadas de divagaciones científicas y filosóficas siempre tan gratificantes.
- A Elena Fuentes, por el apoyo constante y por guiarme en mis primeros pasos del intrincado y a veces bizarro mundo de la economía.
- A Bona M. Vázquez por la confianza y apoyo brindado a lo largo de esta aventura doctoral.
- A Alba Verónica, siempre impulsando las ideas libres y la socialización del conocimiento.
- A mis compañeros y a mis maestros de quienes aprendí que hacer y también lo que no debo hacer.
- A mis alumnos que siempre forman parte de un inmejorable laboratorio para el intercambio de conocimiento.
- A los lectores de este trabajo que con su atinada crítica lo enriquecieron sustancialmente.
- A todos los que fueron directa o indirectamente partícipes de este ciclo de aprendizaje que pretendo sea recurrente.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE CUADROS

CAPÍTULO 1

INTRODUCCIÓN

> "*Claramente, el problema de establecer zonas es relativo al fenómeno que caracteriza al sistema de interés y, por tanto, solo respuestas relativas a este son plausibles*" (Batty, 1978, p 117)

1.1. Planteamiento del problema

Los fenómenos asociados a un contexto socio-económico rara vez cumplen con la condición de uniformidad en el espacio geográfico de referencia donde se presentan. Más aún, este comportamiento no solo ocurre a nivel transversal sino a nivel longitudinal ya que, en la constante dinámica de un sistema complejo de este tipo, el tiempo modifica las condiciones de las unidades espaciales de estudio.

Le definición de regiones entendida como un conjunto de unidades administrativas agregadas es de práctica común en el ámbito socio-económico. Este supuesto asume de manera implícita un distribución uniforme de las unidades espaciales donde se puede presentar un fenómeno de estudio. Se considera también la condición de contigüidad de manera binara, siendo poco realista en cuanto al comportamiento empírico de las variables de estudio el cual se puede presentar de manera difusa en una frontera de índole administrativa.

Bajo el supuesto de uniformidad, el uso de valores agregados por unidad de estudio conlleva a concentrar la información en un valor medio. Sin embargo, en la práctica la intensidad con que se presenta un evento cambia de acuerdo a su ubicación. El valor agregado tomado como medida de una unidad espacial ignora la distribución, en un contexto probabilístico, de la variable de estudio. Tal distribución determina la intensidad con que se presenta el fenómeno asociado a la variable considerando un espacio no uniforme.

El problema se puede resumir como sigue: El uso de agrupaciones a priori como criterio de regionalización puede generar correlaciones espurias o estimaciones sesgadas de funciones de los parámetros asociados a variables de estudio.

1.2. La hipótesis AEC

El fundamento teórico sobre el cual se apoya la metodología propuesta es el denominado "cuarto paradigma" (Hey et al., 2009, loc 70-83). Este concibe a cada unidad de un sistema como un ente generador de información, atributo que identifica a cada unidad por si sola o a las relaciones que se pueden establecer entre estas en términos de flujos o distancias.

El intercambio de información entre unidades no se restringe solamente a un flujo como un proceso dinámico, sino considera también la divergencia que puede existir entre estas una vez que se determinan los factores que las tipifican.

Los factores se representan como variables, las cuales a su vez en su tratamiento empírico se traducen en datos. La intensidad de información obtenida de un sistema se establece mediante la definición de una medida. La obtención de valores específicos de la medida se calcula a partir de un conjunto de datos. Los valores obtenidos son el elemento práctico para la aplicación del método que aquí se propone.

La capacidad actual de recabar y socializar datos posibilita la aplicación del método mediante el uso intensivo en datos, enfoque que se encuadra en la ciencia emergente denominada Ciencia de Datos, alternativa que ha cambiado sustancialmente la manera de

guiar los procesos de investigación en su parte empírica.

El uso de la distribución empírica en lugar de valores agregados en la definición de regiones es relevante dado que un valor de este tipo no representa necesariamente a todo un conjunto, aún inclusive cuando se incluye alguna medida de variablidad. Esto ocurre porque generalmente hay un supuesto implícito acerca de la distribución teórica de probabilidad.

El tratamiento del problema se basará en las primeras dos de las siguientes tres premisas relativas a la configuración espacio-temporal de una regionalización:

- no es libre de contexto;
- está determinada por las variables que representan al fenómeno que la tipifica;
- es dinámica en el tiempo,

De estas premisas y bajo el precepto de que el comportamiento de la configuración de una región establecida por las variables de interés tiene un componente estocástico, se considera la hipótesis denominada *Aleatoreidad Espacial Completa*.

La hipótesis es fundamental ya que establece el principio en el cual se fundamenta la existencia de regiones en un espacio de estudio. El rechazo de la presencia de Aleatoreidad Espacial Completa implica la existencia de grupos o regiones en el universo de referencia o, en forma equivalente, la presencia de patrones en espacios contiguos.

La existencia de *AEC* no significa necesariamente que existe una distribución uniforme de la variable de estudio; esta se refiere a la discrepancia, en términos de la distribución de probabilidad subyacente, que puede existir entre el universo de estudio y las distribuciones de alguna partición de este.

Para validar la hipótesis *AEC* se requiere de un mecanismo que permita establecer una medida de discrepancia entre distribuciones de probabilidad y, más aún, un criterio que establezca si esta discrepancia es o no significativa. Para tal efecto se proponen dos elementos que den respuesta a estos requerimientos:

1. La construcción de una medida de divergencia de información como recurso para establecer la discrepancia entre distribuciones de probabilidad, entendidas estas como patrones de un agrupamiento espacial contiguo;

2. una medida de entropía media como medio para establecer la significancia de la prueba de la existencia de *AEC*, la cual se constituye a la vez como la función objetivo de una regionalización óptima o eficiente.

Estos dos componentes fundamentales del método de regionalización que se propone garantizan dos resultados:

1. Qué la regionalización obtenida será de mínima divergencia de información o, en forma equivalente, la que más informativa es en relación a otras particiones bajo la condición de contigüidad;

2. que una región establecida mediante un criterio de mínima divergencia es, cuando menos, igual de eficiente en términos de la información que aporta comparada con una delimitación establecida a priori.

Probar la hipótesis *AEC* requiere de estos dos resultados. El primero aporta una medida de discrepancia entre las distribuciones empíricas $F_{(n,i)}$ y $F_{(n,j)}$ de las variables en las unidades de estudio correspondientes a los estratos contiguos i y j respectivamente. El segundo resultado dota de un criterio para determinar si se rechaza la presencia de *AEC* con un nivel de significancia dado; además servirá de contraste para comparar una regionalización eficiente con una establecida por un criterio administrativo (a priori).

La validación de la hipótesis se hará a través de tres elementos: un enfoque teórico de información, por construcción de un árbol generado mínimo y, fundamentalmente en su parte empírica, mediante experimentación *in silico*; esto es, a través del uso intensivo de datos a partir de simulaciones y datos reales en el caso específico de las aplicaciones.

El alcance de este trabajo se restringe al caso univariado y a una regionalización en un corte transversal, esto es, no considera el cambio a través del tiempo y, por tanto,

no se da el tratamiento de un proceso estocástico, sino de un comportamiento aleatorio relativo al tiempo en que se realiza la regionalización. Si bien, por la dinámica de variables socio-económicas se puede considerar una estabilidad en cierto periodo de tiempo, este aspecto no es tratado en la investigación.

1.3. Preguntas de investigación

¿Qué tan alejados están, en términos de sus características, los objetos espaciales contiguos geográficamente?

¿En qué medida se distingue un objeto espacial (o conjunto) de todo el sistema (conjunto de objetos) en el que se encuentra inmerso?

¿En qué magnitud difiere una regionalización de otra tomando en cuenta características cuantitativas que las definen?

¿Cuál es la relevancia de considerar la distribución empírica de las características socio-económicas de una unidad espacial como criterio de regionalización?

¿Qué efecto tiene en una regionalización la inclusión de unidades de estudio contiguas externas al espacio de referencia del estudio?

Las respuestas a las preguntas planteadas se abordarán tomando como base la información, en el sentido de Shannon, que cada unidad aporta para una regionalización. Es en este eje sobre el cual gira la discusión y método propuesto para la especificación de regiones.

Particularmente la respuesta a la última pregunta se hará específicamente con la aplicación de regionalización basada en el índice de marginación del estado de Coahuila.

1.4. Objetivos

Objetivo general

Sustentado en el principio de máxima información de una partición, establecer un

criterio general para segmentar en regiones un espacio socio-económico basado en las variables que la tipifican.

Objetivos particulares

- Utilizar el concepto de entropía para definir una distancia entre unidades espaciales de estudio mediante la divergencia de información de distribuciones empíricas;
- definir regiones de máxima información (mínima divergencia) como isomorfismo de la partición de un árbol generado mínimo bajo la condición de contiguüidad geográfica;
- Aplicar la metodología propuesta como criterio para definición de regiones asociadas al índice de marginación y la distribución de ingreso en el estado de Coahuila

1.5. Metas

- Dada la resolución de una partición, crear un algoritmo y desarrollar un programa en lenguaje *R* para la definición de regiones óptimas en el sentido de mínima divergencia.
- Regionalizar, con municipios de Coahuila, tomando como base el índice de marginación y la distribución de ingreso a partir de datos obtenidos en el censo 2010 de *INEGI*.

1.6. Justificación

Si bien la delimitación geográfica (usualmente asociada a un criterio administrativo) considerada como una alternativa de región aporta cierta información del contexto de estudio, este criterio suele conducir a resultados sesgados cuando esta delimitación se refiere a un fenómeno tipificado por una serie de características medibles.

En este sentido, contar con una medida ligada al contexto aporta los elementos necesarios para una delimitación en un espacio geográfico continuo, esto sin eliminar la condición de contigüidad.

El criterio basado en mínima disimilitud en términos de divergencia de información se constituye como un método general que, dependiendo del objetivo de la regionalización, se incorporan las características pertinentes al vector de las unidades espaciales de estudio. Por tanto, es importante establecer que la delimitación de regiones cambiará en función de la estructura dada del vector de variables a utilizar.

El método propuesto rescata la máxima información posible de los datos que se tengan disponibles y su eficiencia pueda ser validada mediante un criterio estadístico en un enfoque empírico.

CAPÍTULO 2

MARCO TEÓRICO

Los conceptos centrales en torno a los cuales se construye este trabajo de investigación son: región, regionalismo, evolución, entropía e información.

La región y el regionalismo son conceptos concomitantes de referencia para la conformación, en este caso, de espacios con carácter preponderante socio-económico. Por un lado las delimitaciones espaciales y sociales se dan en forma natural y se denominan regiones y, por otro lado, la tendencia del hombre a establecer estos espacios se refiere al regionalismo. Por tanto, estos dos elementos coexisten sin necesariamente argumentar una causalidad, aunque si intencionalidad en la definición de algunos espacios, referidos estos en un sentido general[1].

Una visión evolutiva en sistemas sociales y particularmente en los económicos es la contraparte al enfoque mecanicista que, aunque sigue vigente, se ve limitado para explicar el funcionamiento de este tipo de sistemas cuyo atributo principal es la complejidad (Mitchell, 2009). La constante adaptación de elementos e interacciones en una dinámica socioeconómica hacen del enfoque evolutivo una perspectiva robusta para interpretar a este tipo de sistemas; más aún, el principio de *irreversibilidad transaccional* (Ayres, 1994,

[1]La delimitación no tiene que ser geográfica o eminentemente física, ya que hay delimitaciones más abstractas que también pueden considerarse regiones.

p 11, 164, 169) refuerza la adopción de esta forma de concebir un sistema económico[2].

La entropía, vista como la configuración en distintos niveles de excitación de las unidades atómicas de un sistema, tiene su equivalente en la vertiente de teoría de la información, una de las herramientas más utilizadas en la actualidad para aportar una medida de la estructura y dinámica de sistemas complejos. Si bien la interpretación de la entropía en el campo de la física solo es una analogía en un contexto socioeconómico, como medida información toma sentido en esta área del conocimiento.

Los conceptos tratados se dividen en tres apartados: el primero cubre a la región y el regionalismo; el segundo trata los enfoques mecanicista y evolutivo de sistemas sociales y económicos; el tercero se refiere a la entropía desde la perspectiva de termodinámica y su intersección en teoría de la información.

2.1. Región y regionalismo

Para hablar de regionalización primero habría que establecer el significado de lo que es una región, el cual puede ser muy amplio y por tanto se hace necesario acotarlo.

El concepto de región no es nuevo, ha sido ampliamente utilizado por historiadores y también para designar una delimitación de espacios geográficos con características específicas, y es precisamente el sentido geográfico el enfoque más arraigado en la historia del concepto de región (Gasca, 2009, p. 33). En un principio la definición de áreas geográficas no era de interés desde el punto de vista social, sin embargo, el concepto de región, como una delimitación física, evolucionó para darle un sentido de región a un espacio interacción humana. A principios de la década de los ochentas el concepto de región cobró un nuevo significado al incorporar la noción social humana por un grupo de economistas políticos, sociólogos y geógrafos (Storper, 1997, p 3), lo que significó ampliar su dimensión al darle una connotación más allá del sentido geográfico.

[2]Expresado en términos generales, para Kummel (2011, p 176), un sistema económico se conforma por una base física que produce bienes y servicios y una superestructura de mercado donde los actores económicos negocian los productos.

El establecimiento de regiones en el ámbito de la economía, desde el punto de vista de recursos relacionales, destaca la denominada *divina trinidad* de la economía regional, compuesta por tres elementos sustanciales que interactúan entre si: territorio, organización y tecnología (Storper, 1997, p 26). Como un espacio de mercado y producción en masa, los modelos tradicionales se basan en relaciones de intercambio para establecer regiones económicas, sin embargo, la intensa actividad y dinamismo, detonado por la globalización, rompe los esquemas de la forma en que se establece el concepto mismo de región (Vickerman, 2007, p 36). Desde la perspectiva de mercado el establecimiento de región se asocia a mecanismos de producción en masa.

En Europa la noción de región se utilizó para referirse a delimitaciones territoriales de diversa índole, especialmente en la formación de estados y a nivel local para homologar estructuras (Gasca, 2009, p. 34). Conforme se extendió el significado de región se establecieron características particulares que sirvieron como criterio para su determinación. En este sentido, la identificación de una región se enfoca en dos aspectos centrales de contraste, a saber, las especificidades y diferencias entre entornos (geográfico, social, económico, étnico, cultural y político).

Storper (1997, p 3) destaca tres enfoques o escuelas que se integran al debate para establecer criterios de definición de regiones: los que consideran a las instituciones como elemento central, aquellos que se enfocan en las organizaciones industriales y transacciones, y los que toman como elementos principales al cambio tecnológico y el aprendizaje. Cada una de estas propone elementos que justifican la integración del concepto de región como un elemento esencial, no meramente circunstancial o fortuito, de coordinación económica en un entorno eminentemente capitalista (Storper, 1997, p 4).

La *Escuela Italiana* a mediados de los setentas denominada la *Tercera Italia* se caracteriza por los sistemas industriales presentes en centro y noreste de Italia. Piore y Sabel en 1984 intentan capturar este conglomerado en un modelo centrado en la flexibilidad y especialización (Storper, 1997, p 5). Este esquema de organización industrial se toma como una analogía a los *Distritos Industriales* representados en la noción expresada por Alfred Marshall de una *atmósfera industrial* sentada en dos elementos principales, a saber,

la condición de competencia perfecta y la identificación de especificidades de los procesos económicos en este entorno en particular.

Se destacan cuatro aportaciones del pensamiento de la *Escuela Italiana*:

- las tecnologías de producción y división del trabajo no están determinadas por lograr un desempeño óptimo de acuerdo a la dinámica global, sino por las instituciones y decisiones de mercado;
- se identifican la flexibilidad y especialización como catalizadores de la producción en masa;
- las fuerzas dinámicas del capitalismo y las formas avanzadas de aprendizaje tecnológico se encuentran en territorios bien delimitados;
- ante la presencia de incertidumbre las redes institucionales son vitales para la adaptación y persistencia de una región económica.

Las economías externas son el núcleo de la denominada *Escuela de California* centrada en la organización industrial, las transacciones y la aglomeración. El argumento principal se sustenta en la minimización de costos de transacciones como promotor principal del fenómeno de aglomeración, esto realizado a través de un análisis de costos de transacción asociado a encadenamientos interfirmas. Adicionalmente a la flexibilidad y especialización presentada en la *Tercera Italia*, se añade un componente de minimización de riesgo lo que se exige como factor condicionante la proximidad geográfica.

La Teoría del mercado internacional revela una intersección ente el pensamiento de la *Escuela de California* y la *Nueva Geografía Económica* (Krugman, 1991). La concentración geográfica de la actividad productiva en ambas es explicada por los retornos a escala donde prevalece un mecanismo de competencia imperfecta encaminada al dominio de mercado.

Finalmente, la innovación y tecnología son los elementos que establecen un detonador del desarrollo regional, donde ciertas ciertas regiones presentan una mayor propensión al asimilar estos elementos como factor de desarrollo. Este enfoque sería identificado como

la *Escuela Americana*. Al respecto se pueden destacar dos opiniones: la primera encuadra a estas regiones dentro de la escuela institucionalista de especialización flexible, como el caso de Silicon Valley; la segunda visualiza este desarrollo desde el punto de vista de la teoría de la aglomeración y división del trabajo.

La *Escuela Americana* se divide a la vez en dos ramas principales. Una considera esta desarrollo como consecuencia de una alta calidad de vida, buena infraestructura e inclusive buen clima. Otra establece que esto ocurre debido a una política regional, tal es el caso de Silicon Valley por su conexión con la universidad de Stanford y el complejo militar industrial en la zona.

La contraparte europea a la *Escuela Americana* es la desarrollada por el grupo GREMI[3] conformado principalmente por economistas de Francia, Italia y Suiza. Este mecanismo de desarrollo se asocia con lo que Granovetter (1973, p 1360), socio-economista estadounidense, considera como un proceso social y económico integrado. En este sentido, los *milieu* son la materialización al conjuntar un sistema institucional, reglas y prácticas cuyo objetivo principal descansa en la innovación. Se destaca aquí un proceso de circularidad ya que no queda claro si la innovación lleva a la formación de un milieu o la presencia de este lleva a la innovación (Storper, 1997, p 17).

Para Rionda R. (2005, p 15) el "concepto de región es una construcción que identifica a un patrón o parámetro de conducta de una variable de interés". Además, Rionda R. (2005, p 18) hace referencia a Francois Perroux y Jacques Boudeville como los representantes de la escuela francesa, distinguiendo lo que es un espacio y una región. Para Perroux un espacio toma en cuenta tres aspectos:

1. Se define con base en un plan.
2. En relación a un campo de fuerzas.
3. Como un agregado homogéneo.

Esta concepción sirve de base a Jacques Boudeville para la definir tres tipos genéricos

[3]Groupement de Recherche Européen sur le Milieux Innovateurs.

de región.

La conformación de regiones no se da de una manera pragmática y fortuita, sino que obedece a una tendencia humana más profunda denominada *regionalismo*. El concepto de regionalismo es el resultado de la identificación hacia un espacio o territorio de grupos sociales o comunidades, así como de la apropiación de una conciencia cultural o política (Gasca, 2009, p 45). Esto es asociado a una determinada concepción de lo que es una región al establecer elementos que determinen patrones o regularidades persistentes que tipifican el desarrollo de estas delimitaciones (Storper, 1997, p 59), donde no solo el espacio físico toma relevancia sino también el temporal sustentado sobre una base de interacciones.

El establecimiento de patrones que determinan la conformación de una región es fundamental, sin embargo, dicho solo de esta manera se tiene una visión acotada y en cierta manera rígida de identificar cuando una región emerge. Bajo un enfoque evolucionista se asevera que en la presencia de tecnología, en un sentido amplio, no es factible la determinación de patrones o ciclos de desarrollo, al ser la región un lugar donde la proximidad contagia el conocimiento (Storper, 1997, p 80). Con esta visión la región no se concibe como estática en su proceso de formación, pues evoluciona en el tiempo, lo que plantea una nueva manera de visualizar el problema, soportada en una dinámica de interacciones sociales, donde los elementos que determinan a una región no es algo estático, sino se encuentra en constante transformación.

El reto para el análisis de una realidad dinámica en la definición de regiones estriba en identificar aquellos elementos constantes que las determinan. Esta concepción con un componente de incertidumbre se presenta como una analogía al principio de indeterminación de Heisenberg (Georgescu-Roegen, 1999, p 124), donde solo se identifican instantes[4] en un proceso continuo.

Aunque la definición de regiones está asociada al reconocimiento de ciertos patrones, especialistas en sistemas complejos consideran la posibilidad de la presencia, en economía regional, del llamado *Efecto de la Reina Roja*, relacionado con la imposibilidad de predecir el comportamiento en cambios asociados a sistemas dinámicos (Cooper et al., 2010, p V). Más

[4]En el contexto de la economía estos instantes pueden referirse a periodos extensos de tiempo.

aún, la generalización del concepto de región, especialmente en lo que a su delimitación se refiere, toma un connotación difusa como lo menciona Rudy (2005, p 20) al analizar el caso del Valle Imperial de California, identificando distintos factores que le dan forma a este espacio.

Si bien la formación de regiones es fundamentalmente guiada por elementos endógenos, la presencia de unidades económicas organizadas en forma jerárquica se asocia a un efecto antrópico, esto es, de la imposición de formas de organización eminentemente administrativas como lo son localidades, municipios o entidades. Este hecho es uno de los elementos cruciales que motivan este trabajo, ya que este tipo de interferencia es la que genera sesgos cuando se tipifica una región de acuerdo a sus características económicas.

2.2. Mecanicismo y evolución en economía

Distintas visiones de la construcción y funcionamiento de una sociedad se han tenido a través del tiempo. Este devenir se basa en diversas corrientes de pensamiento sustentadas por mentes que han hecho posible el ensamble de este andamiaje de conocimiento. Bajo la premisa de una conciencia universal a escala humana, vertientes equivalentes emergen en forma simultánea y sería imposible mencionar a todos aquellos que contribuyeron en este proceso de continuo cambio. Por tal razón, se destacan aquellos pensadores que se consideran distintivos en el tránsito que va desde una perspectiva mecanicista hasta una visión evolutiva de la sociedad.

En la figura 2.1 se observa la línea de tiempo transcurrida desde 1561 con el nacimiento de Sir. Francis Bacon hasta 2014 tomando como referencia a Robert Ayres y Luciano Floridi.

Aunque no necesariamente se atribuye una influencia explícita de Francis Bacon en la visión mecanicista, este naturalista, filósofo y político tuvo como secretario a Thomas Hobbes de 20 años de edad, una de las mentes más notables y representativas de esta enfoque del funcionamiento de una sociedad, del cual toma posteriormente la analogía de *cuerpo natural* y *cuerpo político* (Ball, 2010, p 67). Esta influencia se detona hasta que

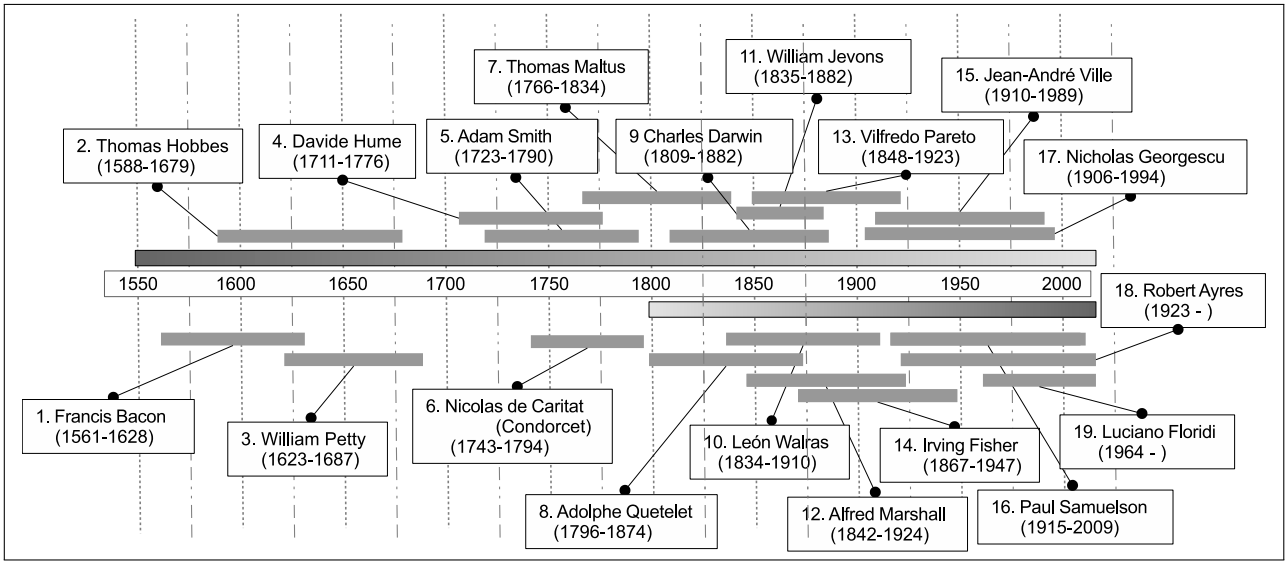


Figura 2.1: *Línea de tiempo mecanicismo y evolucionismo*

Thomas Hobbes tiene contacto con tratados de geometría de René Descartes (Ball, 2010, p 27).

La filosofía mecanicista concibe al universo como un conjunto de piezas que interactúan bajo ciertas reglas bien definidas, interpretación que es trasladada al funcionamiento de la sociedad por Thomas Hobbes y expresada en su obra principal Leviatán, donde emerge como lo que puede considerarse una teoría científica del gobierno (Ball, 2010, p 26). Este pensamiento se enmarca en el denominado materialismo mecanicista donde solo el cuerpo existe más no el alma, dotando al humano de la misma forma de funcionar del universo al heredar las leyes que la rigen.

Thomas Hobbes plantea una explicación reduccionista (Hofstadter, 1982, Mumford y Lill, 2013, p 364-395; 70) de la sociedad, esto mediante la separación en sus partes para ir más allá de entender el mecanismo de su funcionamiento y dilucidar las causas de este. Esta postura es fundamental en su teoría y la contrasta con una visión limitada basada en la cuantificación de una sociedad. Más aún, esta separación en sus partes la lleva hasta un nivel atomicista al considerar al individuo como la parte mínima de una sociedad.

La denominada *Aritmética Política* acuñada por William Petty, discípulo de Thomas Hobbes, considera que la cuantificación es suficiente para darle la categoría de ciencia al

estudio de las sociedades (Ball, 2010, p 14). Es en este punto donde se presenta un contraste entre maestro y discípulo, ya que la crítica de Thomas Hobbes sobre esta concepción se centra en la omisión de las causas que no toma en cuenta William Petty.

Si bien ambos enfoques se encuadran en una teoría mecanicista de la sociedad, Tomas Hobbes va más allá de la limitación impuesta por William Petty al tratar de tipificar una estructura social solo basándose en cifras. Con este avance Hobbes da un paso adelante en dotar a la sociedad no solo de cantidades sino también de causas y efectos.

La intención de Hobbes de explicar las causas del funcionamiento de una sociedad se refiere a los mecanismos de funcionamiento y no a una relación intrínseca causa-efecto de sus componentes. A manera de analogía es como separar todas las partes de un máquina y establecer que parte hereda movimiento a otras partes, estableciendo con esto todas las relaciones que detonan un movimiento total expresado desde una perspectiva holística (Hofstadter, 1999). Con la búsqueda de la dinámica de las partes que conforman una sociedad se plantea un modelo de *física de la sociedad*.

Georgescu-Roegen (1999, p 40) ilustra la pugna entre estas dos versiones de un enfoque mecanicista al destacar dos posturas contrarias: la primera se refiere a que *todas las ciencias deberían imitar a la mecánica*, mientras que la segunda asevera que *sin teoría no hay ciencia*. Estas premisas contrarias enfatizan la diferencia entre lo que denomina *aritmomorfismo* y lo que es ciencia.

La sentencia de Hobbes de que *el valor del hombre radica en su precio* lleva a otra equiparable en significado donde asevera que *la ética del libre mercado es acabar con la competencia*. En concordancia con esta interpretación Ayres (1994, p 134) describe al mercado como un mecanismo colectivo balanceado pero impersonal. Esta concepción de sociedad y mercado llega a tener una gran influencia en Adam Smith (Ball, 2010, p 29). Adam Simith tiene contacto con naturalistas y fisiócratas, destacando a Francois Quesnay[5] (Goodwin, 2012, p 19), reforzando el efecto de la obra de Hobbes en su persona y concebir la idea de un orden social espontáneo, postura que se ve plasmada en su obra

[5]Francois Quesnay es quien, con su propuesta, acuña la expresión "laissez-faire".

una investigación sobre la naturaleza y causas de la riqueza de las naciones[6] (Lüchinger, 2007, p 23).

La aritmética política de William Petty tiene su base en un empirismo puro basado exclusivamente en cifras al que denomina ciencia política. Petty tiene un soporte en las estadísticas sociales, particularmente en las tablas de mortalidad elaboradas por Graunt. Con esta concentración de datos bien organizados se vislumbra el concepto de probabilidad, donde se especula que los datos no revelan lo que ocurre sino más bien lo que pudiera ocurrir (Ball, 2010, p 67). También Petty ejerce una gran influencia sobre Adam Smith (Lüchinger, 2007, p 18), lo cual es razonable si se toma en cuenta la convergencia en gran parte con los postulados de Hobbes. Aunque Hume no comulga con la teoría política de Hobbes por atentar contra la ética y promover la tiranía (Ball, 2010, p 37), ambos mantienen una fuerte influencia sobre Smith, particularmente de Hume esa influencia se refleja en sus teorías éticas y económicas (Lüchinger, 2007, p 25).

El funcionamiento de una sociedad está estrechamente relacionado con la dinámica de crecimiento de su población, situación que hace manifiesta *Thomas Malthus* en su publicación *Ensayo sobre el principio de la población*. En esta obra se enfatiza la importancia de cuantificar, práctica a la que William Petty eleva a nivel de ciencia. Esta revelación de Maltus tuvo gran influencia en Darwin y Marx (Ball, 2010, p 66) ya que Malthus va más allá de esta interpretación general al sentar las bases que conectan con una visión evolutiva, a saber, a través del argumento donde asevera que el crecimiento biológico de la población humana tiene una estrecha relación con los procesos económicos (Georgescu-Roegen, 1999, p 317).

Nicolás de Caritat, más conocido como Nicolás de Condorcet, explora el concepto de probabilidad asociado a juegos plasmando estas ideas en su obra *Esbozo de un cuadro histórico de los progresos del espíritu humano*. Bajo una fuerte influencia del académico francés Jean Le Rond d'Alembert se introduce en asuntos sociales y económicos en el documento cuyo título original es *Essai sur l'application de l'analyse à la probabilité des decisions rendues à la pluralité des voix* (Ball, 2010, p 67-68).

[6]También referido solamente como *La riqueza de las naciones*.

Con un interés similar al de Hobbes centrado en una perspectiva científica de la sociedad, el astrónomo Adolphe Quetelet es el eslabón entre el mecanicismo puro y el enfoque evolutivo de la sociedad. Con una visión ecléctica, conjunta el mecanicismo plantado por Hobbes, la importancia de la cuantificación estadística de Petty y la existencia de leyes naturales de la sociedad al equipararlas con las leyes de Issac Newton. En su artículo *Mecánica Social* muestra como se amalgaman las disciplinas de la física, matemáticas, política y sociología (Ball, 2010, p 67-68).

La aportación más importante de Quetelet es su concepto de *Hombre Promedio*[7] en su mecánica social, donde, a diferencia de otros pensadores, los errores observados en un conjunto de datos no los consideraba como atributos del objeto de estudio sino más bien desviaciones de un comportamiento correcto, revolucionando con esto la interpretación de la estadística como hasta entonces era concebida (Ball, 2010, p 78-79).

EL concepto de evolución en distintos ámbitos ya se vislumbraba aunque de manera implícita con la influencia de los naturalistas. Es con Charles Darwin donde se plantea la evolución en términos de una selección natural mediante un proceso continuo de adaptación al entorno, a diferencia de la teoría Lamarckiana de inducción de características mediante factores exógenos (Ball, 2010, Ayres, 1994, p 86; 17).

La concepción de evolución dentro de un entorno socioeconómico se va dando mediante un proceso iterativo: Darwin recibe la influencia de Maltus, Quetelet conjunta diversos aspectos tratados por Hobbes, Petty y de naturalistas, posteriormente con *El origen de las especies* de Darwin se refuerza la integración de este enfoque en el estudio de la dinámica de sistemas sociales y económicos.

Jevons y Walras sientan los cimientos de la economía moderna, entendida como una extensión del cambio observado en la física, esto al romper con el dogma mecanicista observado también en las ciencias naturales y en la filosofía misma (Georgescu-Roegen, 1999, p 3). Walras es considerado como el padre de la economía neoclásica; desarrolla la teoría de equilibrio de los mercados que finalmente puede ser considerado como el teorema central del neoclasicismo, donde hace un análisis de las formas de curvas de

[7] L'omme moyen

oferta y demanda estableciendo las condiciones de equilibrio estable en su intersección (Lüchinger, 2007, p 67-68). Por su parte Jevons desarrolla con rigor matemático la teoría marginalista, donde plantea que un bien es un elemento abstracto, de tal manera que la utilidad marginal de este se compone de dos dimensiones, a saber, la cantidad del bien que la genera y la intensidad del efecto producido por quien lo posee (BiografiasyVidas, 2000).

Desde Darwin hasta la actualidad se considera que la evolución es necesariamente irreversible, situación que también se plantea en un sistema económico, particularmente porque esto tiene incidencia en el conocimiento humano. A este principio se le denomina *irreversibilidad transaccional* presente en la concepción de Walras del *equilibrio general* (Ayres, 1994, p 11, 164, 169), cuyo principio precursor es la *ley de Say*. Los pioneros de la teoría utilitaria: Jevons, Walras, Pareto, Fisher y Edgeworth, no exploraron las implicaciones de lo que detonaba, a saber, implícitamente sientan las bases de un enfoque evolutivo de la economía.

En su axioma, André de Ville reitera la inexistencia de una ruta que sigan los precios, de tal forma que establezcan una dirección preferida y puedan retornar al punto de partida (Ayres, 1994, p 164). Con este planteamiento se quita la posibilidad de un camino rígido en el ajuste continuo hacia un equilibrio general, situación que se extiende a todo un sistema económico (de mercado).

Shumpeter se opone a la teoría de evolución Darwiniana ya que se basa en un cambio continuo. Por el contrario, Darwin rechazaba los cambios catastróficos. Esto se resume en las dos visiones de cambio radicalmente distintas en un entorno económico: el cambio gradual Usheriano y el súbito Shumpeteriano. Aquí es importante resaltar lo que puede ser la diferencia de fondo: los cambios súbitos no necesariamente implican cambios estructurales sino a condiciones que en el momento requieren ajustes; en contraste, los cambios Usherianos (o Darwinianos) se asocian a cambios estructurales en un sistema socioeconómico (Ayres, 1994, Rosser, 2011, p 154; 121, 126). En el caso del enfoque de Shumpeter este se asocia principalmente a los cambios tecnológicos, lo cual es referido como *innovación redical* (Ayres, 1994, p 155).

Ayres (1994, p XV) denomina *Información útil (SU)* a la información acumulada como un proceso de evolución la cual se caracteriza por incrementos en la cantidad, variabilidad y complejidad de organismos y ecosistemas. Esta tipo de características se observan en los sistemas económicos: incremento en la complejidad estructural; incremento en la importancia de crear, almacenar, procesar y trasmitir información; incremento en la conciencia analítica; incremento en la acumulación de conocimiento. Este último particularmente deriva en el cambio tecnológico, elemento fundamental en la dinámica de un sistema económico.

2.3. Entropía

La entropía es un concepto que se abordará desde dos enfoques principales: en termodinámica entendido como una medida basada en la probabilidad de que cada elemento de un sistema se encuentre en cierto estado y, desde la perspectiva de *Shannon*, como la cantidad de información contenida en un conjunto relacionado con una variable aleatoria. Ambos enfoques pueden ser complementarios cuando se asocian a una economía, ya sea esta vista como un sistema de elementos que interactúan o bien tomando en cuenta la estructura de la misma.

Georgescu-Roegen (1999, p xiii) es uno de los autores que más han incursionado en la equivalencia de leyes de la física aplicados en entornos sociales y económicos, particularmente aquellas relacionadas con la termodinámica a través del concepto de entropía que, aunque más enfocado en el ámbito de la termodinámica, no deja de lado la perspectiva de información.

La definición estadística atribuida a Boltzmann (Kummel, 2011, p 126) asocia el número de estados $\Omega(E)$ de un sistema, la constante de Boltzmann k_B y la entropía S de manera que $S = k_B ln\Omega(E)$. Esta expresión equivale a decir que

> *La entropía de un sistema cuya energía total está en el rango entre E y $E + \delta E$ se incrementa con el logaritmo natural del el número $\Omega(E)$ de macroestados que son accesibles al sistema*

La manifestación física de aleatoriedad está vinculada con la entropía de un sistema, y es precisamente dicha relación la que conecta a ambos conceptos con una medida de información. La unidad natural de información en entropía es el *bit* y en computación es la unidad elemental de información, la cual se codifica como 0 o 1 indicando dos posibles estados de un sistema. Wehenkel (2003, p 42) presenta la medida para establecer la cantidad de información de Shannon, la cual se asocia a la cantidad de información contenida en un conjunto Ω de n mensajes en términos de la probabilidad de ser seleccionarlos. La medida queda establecida por la expresión $I(\Omega) = -\sum_{i=1}^{n} P(\omega_i) log P(\omega_i)$ que sirve como base para muchas otras, por ejemplo, una medida asociada a proximidad entre distribuciones de probablilidad es la de Akaike, que utiliza para su construcción al criterio de información de Kullback Leibler (Konishi y Kitagawa, 2008, p 29).

Ambos enfoques de entropía, termodinámica e información, han sido aplicados en la construcción de modelos económicos, inclusive se ha acuñado el término de *termoeconomía* (Kummel, 2011, Saslow, 1999, p 74, 1239) por el uso de modelos de la termodinámica. Cortés y Rubalcava (1984, p 56, 88) presentan al índice de Gini y de Theil como medidas de concentración, donde particularmente el segundo tiene como base teórica la medida entrópica de información propuesta por Shanon (Wehenkel, 2003, p. 42). Una versión más general del índice de Theil incorpora datos agrupados para obtener una medida de desigualdad (Cortés y Rubalcava, 1984, p 175). Esta alternativa de Theil tiene estrecha relación con el concepto de entropía media, criterio con el cual se medirá la eficiencia de una regionalización.

Otras aplicaciones de modelos de la física en economía son los fundamentados en la teoría de gravitación universal, por lo que se denominan modelos gravitatorios (Roy y Thill, 2004, p 340).

2.3.1. **Enfoque de la termodinámica**

La explicación de la economía basada en la maximización del beneficio y el bienestar queda limitada a factores estáticos, ya que su estructura es difícil de explicar con matemáticas simples, es decir, por su complejidad (Mitchell, 2009, p 1609-27) requiere

de incorporar modelos más robustos. El considerar factores estáticos le da un sentido mecanicista a la manera de tratar los fenómenos económicos ya que no se profundiza en los cambios cualitativos y continuos de fondo.

Con la transgresión del umbral mecanicista, iniciada por Jevons y Walras (Georgescu-Roegen, 1999, p 3), se abre un puente que permite establecer un vínculo entre los mecanismos para explicar fenómenos físicos y el comportamiento de la economía. En este sentido la termodinámica aporta, entre otras cosas, una manera de asociar el intercambio de energía con la generación de valor (Kummel, 2011, Georgescu-Roegen, 1999, p. 172, 276).

La segunda ley de la termodinámica puede ser establecida como sigue (Bryant, 2012, p 59):

> *Es imposible construir un sistema que sea capaz de operar en un ciclo, extraer calor de un depósito, y realizar un trabajo equivalente en el entorno que lo rodea.*

En un contexto económico, esta ley puede ser interpretada de la siguiente manera (Bryant, 2012, p 60):

> *Es imposible construir un sistema económico que sea capaz de operar en un ciclo, extraer contenido productivo de cierto entorno, y realizar un trabajo equivalente, en términos de contenido productivo, en el mismo entorno de referencia.*

En ambas versiones se establece que siempre habrá una fuga de energía, interpretada esta de distinta manera en función del contexto en que se aplica. Esto es, una economía es un sistema abierto.

Saslow (1999, p 1239) apunta que "*un incremento en el beneficio Marshaliano (utilidad marginal) es debido al ocio (beneficio Vebleniano) o al incremento de la eficiencia (beneficio Smithiano)*[8]". La eficiencia de un sistema tiene una estrecha relación con entropía y

[8]Según la tesis central de La riqueza de las naciones, la clave del bienestar social está en el crecimiento económico, que se potencia a través de la división del trabajo y la libre competencia.

es precisamente el elemento que permite establecer un vínculo entre un modelo de termodinámica con la economía (Jaynes, 1991, p 2).

La cantidad económica W, denominada riqueza, se expresa como $W = \lambda M + pN$, donde λ y M representan el valor y cantidad de dinero, p y N son precios y bienes respectivamente; en termodinámica, la energía libre F se define como $F = -PV + \mu N$, donde P es presión, V volumen, μ el potencial químico de partículas y N la cantidad de partículas (Saslow, 1999, p 1240). En estas dos expresiones se relaciona riqueza con energía libre, valor con presión, potencial químico de partículas con valor y cantidad de bienes con número de partículas. En términos de entropía lo que en economía es el excedente $\Psi = U + W$, en termodinámica equivale a $TS = E - F$, donde U es la utilidad, E energía y S entropía. Si bien esta última analogía puede ser más bien interpretada desde un enfoque fenomenológico, el referido a una medida de información se usa con frecuencia en el ámbito de la economía.

2.3.2. **Enfoque de teoría de la información**

El cuarto paradigma considera que todo sistema es generador de información, lo que en un sentido de organismos vivos que interactúan se denomina ***inforg*** (Floridi, 2010, loc 294-402). Tanto en forma individual (intro) como las interacciones entre sistemas (inter) se explican mediante flujos o bien distancias basadas en el concepto de información.

El concepto de información es el centro en torno al cual gira la construcción de una métrica que sirva como criterio de medida de eficiencia a posteriori de una regionalización dada. Esta medida serviría para el establecimiento de una región a priori y delimitada ad-hoc en función de un vector de variables de interés previamente establecidas. Más aún, una medida de información está estrechamente relacionada con la concepción de entropía en el sentido estadístico establecida por Boltzmann (Georgescu-Roegen, 1999, p 144).

La dificultad de establecer una medida de información radica precisamente en el amplio espectro de interpretaciones que se puede dar a esta palabra. La primera tarea es acotar su significado de tal manera que pueda ser trasladada del plano teórico al plano operativo

para su aplicación práctica. Antes de dar una definición formal es importante señalar que la información, en su sentido más amplio, tiene un ciclo típico de vida formado por cuatro fases: ocurrencia, trasmisión, procesamiento y administración, y uso (Floridi, 2010, loc 322, p 4).

Una primera aproximación es establecer lo que se entiende por información tomando como base a los datos que la generan, la cual es pertinente con el surgimiento de la denominada ciencia de datos (Janssens, 2015, loc 178). Para tal efecto, Floridi (2010, loc 561, p 21) presenta la siguiente definición general de información:

Definición 1 *σ es una instancia de información, entendida como contenido semántico, si y solo si,*

1. *σ consiste de n datos, $n > 1$;*
2. *los datos están bien formados;*
3. *los datos bien formados tienen algún significado.*

$\square$

Cabe señalar la relevancia y significado de cada una de las partes que conforman la definición 1. La primera asegura que se cuente con más de un dato en el conjunto que genera información, en la segunda se entiende por un dato bien formado como aquel que sigue reglas sintácticas previamente definidas en el sistema de estudio y, finalmente, la tercera condición es donde la semántica sucede, esto es, otorgarle un significado asociado al contexto de referencia. Una discusión profunda acerca de sistemas formales y su significado semántico es desarrollada por Hofstadter (1999).

La condición 3 de la definición 1 puede tomar dos formas generales, una más apegada a un enfoque positivista donde los datos son generados independientemente de un ser consciente, la cual se denomina *información ambiental*. La contraparte se refiere a aquellos datos generados en forma consciente y con un significado específico (Floridi, 2010, loc 572-580, p 22).

La información puede ser entendida o utilizada de tres maneras: en forma semántica como datos, en forma pragmática como conocimiento y en forma técnica como medida de incertidumbre (Ayres, 1994, p 27). La forma semántica es la concepción más profunda donde se atribuye alguna interpretación; la forma pragmática tiene un sentido menos profundo que no conlleva una interpretación; la forma técnica lleva la información a un plano dimensional al cuantificar el grado de desconocimiento del fenómeno de referencia. El tercer enfoque es el que se utilizará como herramienta base para la determinación de regiones bajo ciertos criterios de optimización que serán establecidos.

Ayres (1994, p 27) asocia la información de Shannon a la probabilidad de ubicarse en un estado o resultado específico entre todos los posibles en un universo físico, usualmente acotado. Esta interpretación la denomina *D-información* ya que asocia la magnitud de la información a una medida de especificidad o duda sobre la presencia de cierto suceso.

Para Floridi (2010, loc 592, p 23) la construcción de una medida de información a partir de datos tendrá su fulcro en la ausencia de uniformidad. En esta principio general es precisamente donde se sustenta la definición de una métrica en el sentido de entropía de Shannon. Entonces, el concepto de *dato* como ausencia de uniformidad puede ser aplicado de tres formas: en el mundo real (físico), entre dos estados de un sistema (sentido de cambio) y entre dos símbolos (codificación).

La conexión de información y evolución es planteada por Ayres (1994, p XIII-XV) con base en cuatro elementos: la presencia de un cambio evolutivo inherentemente impredecible (filogenia); el cambio inherentemente predecible (ontogenia); la información útil (SU) para la sobrevivencia; la información relevante (SR) para la sobreviviencia.

Ayres (1994, p XV) destaca la relevancia de la información útil al conectarla con las características de un sistema económico, a saber, que refleja la complejidad estructural, la importancia de almacenar, procesar y trasmitir, la conciencia analítica y especialmente la acumulación de conocimiento como ruta de cambio. Considera que la evolución Darwiniana padece de *miopía* al tomar en cuenta solo óptimos locales, en contraste, desde la perspectiva de información de la evolución, esta presenta *presbiopía* al tener un campo de visión más amplio.

2.4. Anotaciones

La conjunción de los tres apartados conforman el marco teórico que sustenta la propuesta para establecer regiones óptimas en el sentido de información.

El concepto abierto de región y la importancia de formarlas para una mejor comprensión de la dinámica de sistemas económicos plantea la necesidad de desarrollar metodologías que, bajo criterios bien establecidos, deriven en la delimitación, en un sentido amplio, de lo que se considera como región.

El constante cambio y la complejidad de un sistema económico da pie a incorporar otras maneras de entender a sistemas de este tipo. Si bien el enfoque evolutivo en este ámbito no es nuevo, este no se ha utilizado en forma generalizada como una extensión de un esquema mecanicista centrado en comportamientos más rígidos y que se distinguen por sustentarse en modelos estáticos.

La principal herramienta para develar distintas configuraciones de regiones económicas se basa en el concepto de entropía. Este concepto, encuadrado en el ámbito de la teoría de la información, se toma como base para establecer una medida de eficiencia y de pureza al evaluar particiones de un entorno económico que derivan en diversas regionalizaciones.

El sentido dinámico de una topología regional socioeconómica da pie a la definición de regionalizaciones diversas dependiendo de las características atribuidas a priori a un espacio de este tipo. La característica distintiva de este enfoque es el de romper con la condición de un modelo estático, tanto en el espacio como en el tiempo, de una estructura regional.

Si bien la configuración regional a través del tiempo es relevante, este trabajo se enfoca a la estructura regional en un corte transversal definido por el momento en que se obtienen los datos de variables establecidas para, posteriormente, generar la información necesaria y así obtener la delimitación de una partición del espacio de estudio.

CAPÍTULO 3

METODOLOGÍA

Ante la falta de una definición general de región, lo que se tiene estaría estrechamente relacionada con el objeto de estudio, las características del mismo y el contexto en el que se aplica, la primera parte se centra en documentar las interpretaciones de lo que se entiende por región socioeconómica.

Si bien no se cuenta con una definición universal de región, si se identifican mecanismos pragmáticos o escuelas que dan forma a distintos tipos de aglomeración. Partiendo de este precepto, se documenta una clasificación general de agrupaciones de índole socioeconómica de acuerdo a su topología, características, interacción o funcionamiento. Aquí es fundamental considerar la diferencia sustancial entre lo que es una regionalización y lo que es una agrupación. Mientras que la primera considera aspectos estructurales, la segunda usualmente se hace mediante una clasificación de valores agregados o bien de criterios administrativos y geográficos. El concepto de región que se adopta se basa en la condición de homogeneidad en un espacio económico que fue planteado por Perroux (Rionda R., 2005, p 18).

Se hará una compilación de los principales métodos de regionalización para identificar sus características, ventajas y desventajas.

La naturaleza de esta investigación es de índole cuantitativa y basada en el estable-

cimiento de métricas. Por tanto, se hará una revisión de las principales medidas y sus propiedades que son utilizadas en la identificación de regiones, particularmente aquellas que sirven como medida de disimilitud entre grupos, lo cual es base para extender una clasificación a lo que definiremos como región.

Se hace una revisión exhaustiva del concepto de entropía tanto en su enfoque de termodinámica como en el de teoría de información. Este apartado tiene una relevancia especial ya que es en este concepto en el que se centra la propuesta de establecer un método general de regionalización óptima en el sentido de información.

En relación a los métodos de regionalización, se documentan aquellos encaminados a la construcción eficiente de regiones en relación a la métrica utilizada. Se presenta y reproduce el algoritmo para la construcción de árboles generados mínimos (MST) (Assuncão et al., 2006, p 798) el cual será utilizado como base en el establecimiento de regiones de máxima información. En este sentido, se propone una medida de divergencia de información como alternativa de disimilitud. Se utiliza el algoritmo *PRIM* para la construcción del MST y un criterio híbrido para la remoción de aristas.

Para establecer la eficiencia del método de regionalización basado en MST y máxima información en la identificación de grupos afines (clusters), se realizarán simulaciones y contrastes entre los métodos y métricas seleccionados para validar su comportamiento y robustez. La razón de utilizar esta alternativa es que se conoce en forma explícita la naturaleza de la población de estudio, teniendo de esta manera una referencia del nivel de desempeño del método en la detección de grupos mediante el control de parámetros de sensibilidad.

El procedimiento para construcción de regiones, su identificación, la validación de hipótesis y de eficiencia, se realiza mediante lo que se denomina experimentación in silico (Taverna, 2015).

Los datos requeridos para realizar tanto las simulaciones como la aplicación empírica serán organizados de tal forma que cumplan con los requisitos de una estructura de datos. Esta parte implica la homologación de índices de variables clave para poder ser enlazados con diversas fuentes de información como lo es, por ejemplo, los datos generados por

INEGI. Este proceso es fundamental en el trabajo ya que reduce la cantidad potencial de errores al verificar la integridad referencial de los componentes de las bases de datos, además de poder acceder a estos de diversas herramientas particularmente del software para procesamiento estadístico que se utilizará.

En cuanto a la aplicación en distribución de ingreso, se contrastan la medida de información de Kullback-Leibler, el índice de Gini y el ingreso medio para constatar como se obtienen distintas clasificaciones de los municipios de Coahuila. Los microdatos utilizados corresponden al censo 2010 de *INEGI* y es un total de 56,298 personas cuyo ingreso por trabajo[1] es mayor que cero y menor a 999,998.00. No se incluye este valor ya que corresponde a personas que declararon ganar más que esa cantidad, lo que introduciría un sesgo en la estimación de la distribución empírica.

Para el caso de la aplicación relacionada con el índice de marginación se requiere realizar tres tareas principales: la primera es establecer una transformación adecuada para acotarlo en el intervalo de 0 a 1, ya que actualmente como lo calcula CONAPO nominalmente no esta acotado, al menos no de forma equiparable entre distintas mediciones en el tiempo, lo que dificulta su lectura principalmente; la segunda se refiere a hacerlo comparable en el tiempo, esto es, que no se restrinja solo a realizar un ordenamiento transversal de las unidades de estudio, sino que sea factible hacer contrastes longitudinales para ver la evolución del mismo; la tercera es calcularlo directamente de datos de censos con la capacidad de incorporar variables adicionales de acuerdo al contexto de estudio.

Una vez establecida una regionalización del índice de marginación, se analiza la influencia que pueden tener las entidades contigüas a Coahuila. Para tal efecto, se compara el kernel de la primera regionalización con el obtenido si se integran localidades rurales de municipios colindantes a Coahuila de las entidades Chihuahua, Durango, Zacatecas y Nuevo león.

[1]Se detallan en el capítulo de aplicaciones las características de este tipo de ingreso.

3.1. Herramientas utilizadas

Dado que la investigación es de índole teórico y empírico, se fundamenta la aplicación de los métodos aplicados bajo un un enfoque de ciencia de datos. Particularmente se conjuntan herramientas tecnológicas para hacerlo en un esquema de investigación reproducible (Grandrud, 2015, p 3-4). Los principales recursos tecnológicos utilizados son: MySQL (DuBois, 2006), Lenguaje R (Adler, 2009), Octave (Eaton et al., 2008, Quarteroni et al., 2010), QGIS (Sherman, 2012), LaTeX (Diller, 1999) y LibreOffice (Fundation, 2015).

Las bases de datos provenientes de INEGI están elaboradas en un formato dbf, las cuales contienen medidas resumen, haciendo de esto una base redundante. Para limpiar estas bases se eliminan aquellos campos redundantes y se añaden otros necesarios para poder homologar con claves unificadas, particularmente aquellas para identificar las unidades de estudio y establecer vínculos con la estructura en sistemas de información geográfica o propietarios como el caso SCINCE en INEGI.

El nodo para el procesamiento y conexión con los diversos programas utilizados es el lenguaje R. En este lenguaje se realiza toda la programación para el procesamiento de datos, generación de tablas resumen, elaboración de gráficas y su conversión a diversos formatos para su edición y presentación en versión digital e impresión.

Octave se utiliza como apoyo en el procesamiento y simulación enfocada al cómputo numérico con énfasis en al manejo de datos dados en forma matricial y vectorial.

La georepresentación de resultados se realiza en el programa QGIS. Este programa interactúa de manera bidireccional con el lenguaje R. De QGIS se toman las bases fuente dadas en dbf para ser utilizadas como insumo en el lenguaje R; una vez calculados valores de indicadores se anexan nuevos campos a la base original para posteriormente ser representados espacialmente.

En la elaboración del documento impreso se utiliza el lenguaje de etiquetas LaTeX. Este lenguaje está diseñado para la elaboración de documentos científicos de alta calidad y es una de las herramientas esenciales para hacer investigación reproducible. El mecanismo por el cual LaTeX recibe información actualizada es a través del lenguaje R: al procesar

los scripts de R se generan elementos como tablas resumen y gráficos que son ubicados en la estructura LaTeX. Una vez ejecutado un script de R se compila el texto en donde se tendrán los elementos que lo conforman actualizados en forma automática.

En la suite de oficina LibreOffice se elaboran presentaciones, hojas de cálculo y diagramas. También se interactúa a través del lenguaje R particularmente con el módulo *Calc* en forma bidireccional: R lee tablas elaboradas en Calc y puede a la vez retroalimentarlas con resultados actualizados dentro de hojas de cálculo. Este programa es compatible con la suite comercial *Office*.

CAPÍTULO 4

REGIONALIZACIÓN: ESTADO DEL ARTE

Para enfatizar el aspecto que se toma como soporte principal en el método a desarrollar se parte de la siguiente aseveración atribuida a Fortin (2009, p 89).

"*No existe un patrón espacial sin un proceso subyacente que lo genere*"

Como ya se ha mencionado, ante la falta de una definición unificada de región socio-económica, se parte de una tipología de lo que se puede enmarcar como región basándose en ciertos atributos o interacciones. Además, bajo este precepto se describe lo que es un proceso de regionalización, entendido este como la división de un espacio socioeconómico tanto en lo geográfico como en lo referente a los atributos de referencia.

4.1. Tipos de región

La conformación de regiones en muchos de los casos se define por delimitaciones políticas o administrativas; por ejemplo en México los estados son de alguna manera regiones por si mismas (o al menos así son tratadas de forma errónea), o al interior de los estados los municipios replican este patrón y se consideran regiones en un contexto estatal.

Regionalizar se puede considerar como el proceso funcional para establecer una división en estratos de cierto entorno y bajo un contexto de referencia. En este sentido Gasca (2009, p 44) menciona que la *"regionalización es un concepto relativo que está en función del enfoque con que se aborda y se conceptualiza el tipo de región o fenómeno regional tratado"*. Esta concepción es consistente con la dada por Fortin (2009, p 89).

Boudeville, apartir de la definción de un espacio económico establecida por Perroux (Rionda R., 2005, p 18) establece tres tipos genéricos de región: 1) región plan o programa, 2) región polarizada y 3) región homogénea. Estos a la postre forman parte de una clasificación más extensa de tipos de región.

En función de las características o criterios que se establecen para determinar una región se definen siete tipos (Gasca, 2009, p 35-42):

- homogénea;
- nodal o funcional;
- sistémica;
- política;
- plan o programa;
- economía política;
- cultural;

La región homogénea se refiere a una espacio uniforme, lo que permite encontrar diferencias y similitudes con otras regiones, centrándose en determinar la mínima diferencia al interior de cada región; la nodal o funcional se define a través de ámbitos espaciales donde se identifican relaciones funcionales entre elementos que la conforman (Farrell y Héritier, 2005, p 274); la sistémica, es entendida como un conjunto espacial integral, cuyo distintivo son sus relaciones dinámicas en una variedad de aspectos como lo físico, social, cultural y económico, dándole un sentido holístico; la política refiere a delimitaciones territoriales

que asocian estructuras institucionales, con la finalidad de crear organizaciones políticas y administrativas (Paasi, 2004, p 537); la plan o programa se vincula a una intervención del gobierno, particularmente en el ejercicio del presupuesto en aplicación de programas sociales; la economía política se relaciona con la división espacial donde están presentes los conceptos de trabajo, capital y mercados laborales, donde el capital es un elemento central en su definición; la cultural se forma mediante la identificación arquitectura, flujos de conocimiento, de identidad y pertenencia socio-espacial.

Para establecer una región se determinan criterios que la definan; de acuerdo a Donaghy (2010, p 3), en la actualidad los avances metodológicos en la definición de regiones se resume en tres modelos principales: redes, equilibrio general y econométricos, cada uno de ellos con características particulares. Por ejemplo, en un estudio económico se desarrollaron demarcaciones espaciales (denominadas regiones funcionales) basadas en una desagregación jerárquica, procurando delinear mercados de trabajo regionales en los que la geografía resultante tuviera un significado inherente (Mitchell, 2009, p 38).

Siguiendo la metodología de modelo de redes, Jan y Gijsbertus (2010, p 133) desarrollan un modelo de transporte en un marco de escalas espaciales, mientras que Wixted y Cooper (2010, p 183) analizaron la evolución de redes inter-cluster entre nueve economías de la OCDE.

En relación a modelos de equilibrio, Wing y Anderson (2010, p. 263) utilizaron modelos multiregionales para definir pequeñas áreas económicas.

La evolución en economía se refiere a la manera en como la economía selecciona una o varias estructuras geográficas posibles. Además, la presencia de múltiples equilibrios es un distintivo relevante en los modelos de la nueva geografía económica. Esto se asocia con el denominado nuevo regionalismo, derivado del proceso de globalización cuya dinámica en un contexto económico induce una nueva territorialización (Lu, 2011, p 335).

Las regiones formadas como efecto de la globalización no son la única manera de entender al nuevo regionalismo; existe también la tendencia a establecer delimitaciones geográficas ad-hoc al darles coherencia mediante criterios preestablecidos, tal como lo describe Lu (2011, p 335) en la formación de regiones rurales, donde tomó como

elemento de su establecimiento al desarrollo económico. Brenner (2000, p 320) analizó la construcción de regiones en Europa como espacios políticos establecidos bajo una estrategia corporativista contemporánea; aquí el término construcción se asocia al hecho de que hay una intención expresa para establecer regiones.

4.2. Métodos de regionalización

La utilidad de las diversas formas, en un sentido amplio, que se presentan en un contexto espacial depende de la naturaleza intrínseca del objeto de estudio. No considerar esta dependencia genera el denominado *Problema de la Unidad de Área Modificable* (MAUP[1]) planteado por Openshaw (1984, p 3) y referido en un contexto multivariado por Fotheringham y Wong (1991, p 1026). El problema tiene su raíz desde la conceptualización de lo que es una región, donde en la mayoría de los casos no toma en cuenta el objeto de estudio como determinante en la delimitación de esta.

El objeto de estudio no debe ser entendido como una mera entidad física, sino como el conjunto de características[2] que lo definen de tal manera que estas determinen el nivel de agregación requerido, bajo un criterio explícito, para generar conjuntos bien delimitados en el sentido de evitar la ambigüedad al momento de establecer una región. La práctica más común en la definición de regiones se lleva a cabo a través de consideraciones políticas locales o mediante criterios de administración de gobierno, situación que deriva en áreas con diversos significados intrínsecos potenciales.

Los resultados derivados de una regionalización discresional tienen efectos importantes cuando se aplican para establecer predicciones o, más aún, cuando esto es utilizado para le diseño de políticas públicas. Dependiendo del nivel de agregación se pueden tener resultados diferentes significativamente, esto es, la presencia del *MAUP* no puede ser negligible en validación de hipótesis o en toma de decisiones.

La división de un espacio económico puede ser realizado de múltiples maneras ya que

[1]Modifiable Aereal Unity Problem.

[2]Avatares abstractos dados como un vector de características que tipifican al objeto de estudio.

está ligado al contexto de las características seleccionadas que definen al objeto de estudio. Un vez establecido un vector de características se requiere un criterio para determinar, de todas las posibles particiones, cual es la que representa un óptimo de tal manera que se tenga una regionalización eficiente de acuerdo a dicho criterio. Lo deseable es que los métodos de regionalización posean métricas que permitan jerarquizar el grado de eficiencia de cada elección realizada.

Masser y Brown (1978a, p 1) describen el problema que se presenta cuando se trata con la agregación de variables continuas para formar grupos considerados como unidades discretas. En el proceso de construcción de grupos homogéneos se pueden tomar en consideración variables asociadas a propiedades inherentes de las unidades de estudio o bien aquellas que representan interacciones entre estas. Estos dos enfoques plantean retos específicos en la conformación de estratos y particularmente en la interpretación de cada configuración obtenida.

En este sentido se distinguen dos problemas específicos: el problema de agregación multi-criterio (MCA[3]) y el de espicificación multi-nivel (MLS[4])

En el proceso de división en regiones de un espacio que contiene cierta cantidad de zonas presenta dos problemas: el de escala y el de agregación Openshaw (1984, p 8). El primero se refiere a la posibilidad de agregar distintas cantidades de unidades de estudio; el segundo es considerar una cantidad fija de unidades. Ambos tendrán efectos significativos para la elección de distintas regionalizaciones. Estos problemas se manifiestan de distinta manera, particularmente en la cantidad de regiones potenciales, si se tiene la restricción de contigüidad de zonas o se relaja para tener regiones no conexas que denominaremos *dispersas*. Los sistemas y problemas de regionalización se representan en la figura 4.1.

Assuncão et al. (2006, p 797) establecen que:

Definición 2 *La* regionalización *es una proceso de clasificación aplicado a objetos espaciales con una representación de área agrupados en espacios homogéneos continuos.*

[3]Multi-Cirteria Aggregation.

[4]Multi-Level Specification.

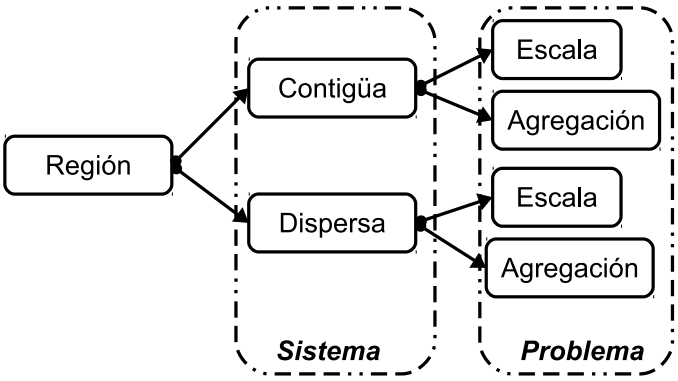


Figura 4.1: *Problemas y sistemas de regionalización*

⊡

La definición 2 presenta un panorama amplio de lo que es la regionalización donde la continuidad se entiende en este caso como contigüidad. Aquí se presentan tres elementos sustanciales en una regionalización: agrupación, homogeneidad y continuidad. Si bien se dan los componentes esenciales aun no se aporta el mecanismo con que se realiza dicho proceso. Además, dependiendo del contexto en que se aplique esto podrá derivar en diversas estructuras.

En general existen tres técnicas de regionalización: la primera se realiza en dos etapas que consisten en un algoritmo de clusters no espaciales seguida de una clasificación preservando la vecindad de los elementos; la segunda utiliza coordenadas geográficas como atributo adicional en el procedimiento de clusters; en la tercera la relación de vecindad entre los objetos espaciales es utilizada de manera explícita en un procedimiento de optimización (Assuncão et al., 2006, p 798) al considerar distancias basadas en las características de los objetos espaciales.

La idea central de esta tesis se sustenta en la consideración de que no se pueden disociar las características de los objetos espaciales en la delimitación de una región. Por tanto, la tercera de las técnicas de regionalización es la que se desarrollará mediante la inclusión de métricas alternativas.

Openshaw (1984, p 3) establece un procedimiento desarrollado en forma exhausti-

va para construir una regionalización donde se minimiza una función objetivo. A este procedimiento se le denomina *Procedimiento de Zonas Automático* (AZP por sus siglas en inglés) y se realiza mediante ensayo error al considerar distintas agrupaciones hasta lograr el óptimo. La observación inmediata es que hecho de esta manera resulta costoso computacionalmente, por lo que es deseable contar con un procedimiento que sistemáticamente genere una regionalización óptima de acuerdo al criterio establecido por la función objetivo.

La alternativa que presenta (Assuncão et al., 2006, p 799–806) utiliza la construcción de árboles a partir de grafos. Al mecanismo para transformar un problema de regionalización en uno de partición de grafos se conoce como *Análisis de Clusters Espaciales mediante Remoción de Aristas de un Árbol* (SKATER[5]). Los pasos se describen enseguida:

1. se parte de un grafo conexo de tal manera que se crean aristas entre los centroides de los objetos espaciales considerados como unidad mínima a ser agregada en regiones. A cada arista se le asigna una medida de disimilitud expresada en términos de las características de cada unidad espacial;

2. la complejidad del grafo inicial se reduce eliminando aristas con más alta disimilitud hasta generar un *Árbol Generado Mínimo* (MST[6]). El criterio utilizado para definir un *MST* es a través de normas euclideanas como medida de disimilitud.

3. una vez que se tiene un árbol generado mínimo cada arista eliminada genera subgrafos candidatos a formar un cluster. En esta fase se requiere de un método específico donde se establezcan las restricciones para el nivel de resolución de la subdivisión del *MST*.

La esquematización de las tres fases se muestra en figura 4.2.

Cada una de las fases para generar particiones del *MST* requiere un método en si mismo. En el caso de la primera la condición establecida es la contigüidad. En el caso

[5]Spatial Kluster Analysis by Tree Edge Removal.
[6]Minimum Spanning Tree.

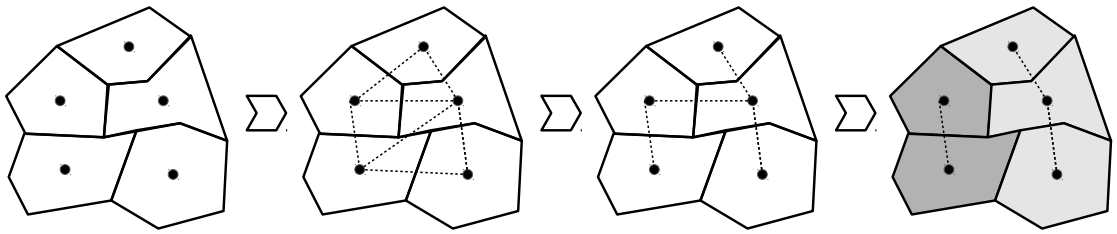


Figura 4.2: *Proceso para generar clusters en un árbol mínimo*

de las dos siguientes un elemento central para el resto del proceso es la definición de la métrica para determinar la disimilitud, ya que esta establece el criterio de afinidad entre objetos espaciales contigüos y por tanto también el resultado final de una regionalización. El proceso de construcción de regiones óptimas se desarrolla con detalle en el capítulo 5.

CAPÍTULO 5

INFORMACIÓN, ENTROPÍA Y REGIONALIZACIÓN

Está en la naturaleza humana buscar patrones cuando se encuentra inmerso en un entorno complejo que genera información (Waller, 2009, p 299). Estos patrones pueden ser de distinto tipo y generan aglutinaciones definidas por las características inherentes al objeto de estudio; estas aglutinaciones o grupos son referidos como clusters.

En la identificación de grupos la palabra clave es *información*, concepto que se abordará principalmente desde la perspectiva de Shannon y servirá como criterio en la definición de una métrica de disimilitud entre unidades espaciales de estudio.

Dentro del contexto de termodinámica, en forma paralela al concepto de información, se tiene el de *Entropía* en el sentido de *Boltzmann*, cuya interpretación en términos de orden y distribución de energía (Kummel, 2011, p 29, 113) juega un rol central en la descripción de un comportamiento dinámico de sistemas complejos como lo es la economía.

Tanto la información como la entropía juegan un papel preponderante para establecer los criterios en la conformación de regiones socioeconómicas. Por tanto, son estos conceptos los que darán soporte a la metodología propuesta para la definición de regiones espaciales óptimas bajo un criterio de contigüidad espacial en un entorno socioeconómico.

5.1. Regionalización eficiente

La formación de regiones en un sentido amplio no se restringe a su concepción como agrupaciones físicas (geográficos) con la condición de contigüidad; esto es, se debe distinguir entre lo que es una *agrupación* y una *regionalización*. Una región también tiene un sentido abstracto, esto es, definido por la relación entre características que presentan las unidades de estudio. De hecho, las relaciones entre variables socio-económicas es lo que en esencia conforma a la econometría donde el principal objetivo es explicar de manera cuantitativa las asociaciones entre características expresadas como variables. Otro aspecto relevante es que las asociaciones entre atributos pueden cambiar en el tiempo tanto en intensidad como en dirección, lo que refuerza en tal caso la constitución de regiones dinámicas y el enfoque evolutivo en su tratamiento.

Ante las limitaciones que representa trabajar con datos y la importancia que tiene en un contexto de políticas públicas, este trabajo considera en primera instancia unidades administrativas como objeto de estudio bajo el criterio de contigüidad pero bajo el concepto de región dinámica en un sentido evolutivo. Esto es, que aunque se mantenga la contigüidad, la configuración regional cambiará con el contexto y en el tiempo, esto sin menoscabo del objetivo central que es el de identificar similitudes entre las unidades de estudio.

Sin embargo, dada la condición difusa de una frontera administrativa expresada en función de las características del objeto de estudio, es factible delimitar una región mediante la formación de clusters de subunidades que la conforman, por ejemplo considerar a un municipio como la colección de localidades o una zona urbana como el conjunto formado por *AGEBs*.

La conformación de regiones mediante fronteras administrativas es de índole holista siendo esta la más utilizada. Desde la perspectiva de un conjunto como colección de sus elementos es considerada con un enfoque reduccionista. La segunda manera de concebir una regionalización es más realista desde la perspectiva de la afinidad de características entre objetos ya que reduce el sesgo inducido por una frontera establecida a priori.

5.1.1. **Identificación de regiones**

El establecimiento de regiones se reduce a la identificación de grupos o clusters mediante la presencia o intensidad de ciertas características comunes. La presencia se asocia a variables discretas mientras que la intensidad a variables continuas definidas por los atributos de las unidades espaciales de estudio. En general, los métodos se centran en la definición de una medida de disimilitud utilizada para contrastar lo observado en cierta población bajo un esquema de hipótesis (Waller, 2009, p 303) general planteada como sigue:

$$\mathcal{H}_0 : \text{ No existe presencia de clusters}$$
$$\mathcal{H}_1 : \text{ Existe presencia de clusters}$$

Una situación común es considerar que las unidades espaciales donde puede ocurrir un evento se asumen uniformemente distribuidas y la cantidad proporcional al área que cubren. Este supuesto no siempre es válido y es necesario diferenciar entre una ubicación donde un evento *puede ocurrir* y donde un evento *ocurre*. Mientras que el primero se refiere a la distribución espacial de las localidades el segundo se asocia a la distribución espacial de la localidad donde el fenómeno de estudio sucede. La hipótesis asociada al primero de los casos se establece como sigue:

$$\mathcal{H}_0 : \text{ Aleatoreidad espacial completa (AEC)}$$
$$\mathcal{H}_1 : \text{ No existe aleatoreidad espacial completa}$$

En notación formal de hipótesis estadística se expresa como:

$$\mathcal{H}_0 : \text{AEC} \tag{5.1}$$
$$\mathcal{H}_1 : \sim \text{AEC}$$

Bajo la hipótesis nula (*AEC*), la presencia de un evento de estudio es proporcional a su área y definida por una constante λ. Por tanto, la cantidad de eventos esperados en una unidad con área $|A|$ sigue una distribución de Poisson con parámetro $\lambda|A|$. Además, dada una cantidad de eventos, en una primera instancia se considera que estos están distribuidos de manera uniforme en el área de referencia.

Aunque el modelo nulo *AEC* es utilizado con frecuencia en la literatura se debe considerar el caso en que la población en unidades espaciales puede no estar distribuida en forma homogénea. Una manera de lidiar con esta condición es mediante la definición de una función $\lambda(x)$ como el valor esperado de eventos por unidad de área en el punto x. A la función $\lambda(x)$ se le denomina función de intensidad, dando como resultado un proceso de Poisson heterogéneo ya que se pondera por esta medida de intensidad (Fotheringham et al., 2000, Waller, 2009, p 304; 145).

Una vez que se tiene establecida la mecánica para establecer la cantidad de eventos en unidades espaciales se analizan las características similares, considerando en este caso la condición de contigüidad. Esta condición se utiliza dada la tendencia de que ciertas unidades tengan características similares a las más próximas y para la toma de decisiones en política pública, de tal manera que el problema se reduce a identificar a los que se consideraría vecinos (Waller, 2009, p 306, 307) cuya raíz está en la denominada *autocorrelación espacial* (O'Sullivan y Unwin, 2010, p 34).

La autocorrelación espacial toma relevancia en un contexto eminentemente geográfico, sin embargo, la distancia utilizada no se reduce a una métrica física, sino que se extiende a una distancia en términos de las características de las unidades espaciales. Por tanto, la definición de cluster que se pretende establecer considera la condición de contigüidad pero utilizando la similitud (disimilitud) en los vectores de características que definen a

cada una de estas. Esto es fundamental ya que bajo este criterio se establecerá, para cada unidad, cuales se consideran vecinas de estas.

5.1.2. **Región dinámica y el enfoque evolutivo**

La construcción de una región óptima se asocia al problema de maximizar, utilizando alguna métrica, una o más funciones objetivo que están estrechamente relacionadas con el objeto espacial de estudio. La relación se establece conforme a un conjunto de características que tipifique a las unidades de estudio para que la regionalización obtenida sea consistente con la estructura de los elementos que la conforman. Es en esta definición de características donde se establece el contexto que determinará la configuración regional óptima.

Tomando como base el enfoque evolutivo, una configuración regional dependerá del contexto y el momento de su definición. Esto da sustento al precepto de región dinámica ya que esta puede cambiar debido a las dos condiciones planteadas: por la selección del fenómeno que la define (contexto) y por el momento en que se define (tiempo). Esto significa que una regionalización establecida es válida en ese corte transversal, sin embargo a gran escala en términos de la cantidad de unidades de estudio, se esperaría que se requiera de un tiempo considerable para observar cambios en topología regional.

5.2. Detección de clusters

Dado un vector de características que define a las unidades espaciales, se requieren como base dos elementos para establecer subconjuntos denominados clusters: una medida de proximidad y un criterio de agrupación. La proximidad se determina mediante la definición de una distancia, esto es, se debe de establecer la métrica a utilizar; la formación de estratos requiere de un criterio de agrupación, el cual en buena medida está determinado por el contexto al que se circunscriben las unidades espaciales.

Entre las distancias más utilizadas se destacan la euclideana, máximo, manhattan,

canberra, binaria y minkowsky (apéndice A). Estas distancias se establecen en relación a una geodésica, esto es, un distancia geográfica definida en cierta superficie o recorrido. En conjunción con estas, una distancia en términos de maximización de información, como lo es la de *Kullback-Leibler*, se relaciona con una distancia estructural (distribución empírica de probabilidad) de las unidades espaciales. La combinación de estas métricas es la que da soporte al método propuesto para la definición de regiones óptimas.

Los conceptos de grafo, subgrafo, árbol y árbol generado mínimo en un contexto geométrico son isomorfos a diferentes instancias en el proceso de formación de grupos de unidades espaciales afines en un contexto socio-económico. Para la definición de regiones estos son determinantes tanto en su representación como en la metodología a utilizar. Las definiciones formales de estos conceptos se ven con detalle en la sección 5.3.

Openshaw (1984, p 22) utiliza como función a maximizar el coeficiente de correlación entre variables de estudio y una metodología de clusters (Jobson, 1991, p 483) para formar los grupos. Assuncão et al. (2006, p 799) define a un objeto como un vector de características y su disimilitud se basa en la norma euclideana, todo esto realizado a través de la definición de árboles, subgrafos y minimización por técnica de clusters. Al vector formado por las características de las unidades espaciales también se le da el nombre de ávatar de características o ávatar electrónico[1].

El proceso para generar subgrafos (clusters) en un *Árbol Generado Mínimo* (*MST*) requiere de tres etapas: la definición, la construcción y la partición. En la primera etapa establece la notación y métrica utilizada para la dismilitud que conforma la malla (grafo) inicial determinada por la condición de contigüidad. La segunda se refiere al algoritmo específico para eliminar aristas de máxima disimilitud para obtener un solo árbol. La tercera parte se enfoca al algoritmo utilizado en la remoción de aristas para crear un conjunto de subgrafos que definen la regionalización óptima de acuerdo a criterios que se establecerán.

[1]Por ejemplo la estimación de parámetros de salud en un contexto médico (Hey et al., 2009, loc 2137).

5.2.1. Intensidad espacial

Una de las vías para lidiar con la falta de uniformidad en la distribución espacial de eventos es mediante la ponderación basada en una función de intensidad, denotada por $\lambda(x)$. Una de las formas más comunes de estimar esta función es mediante el uso de otra función denominada kernel (Bivand et al., 2008, Waller, 2009, loc 2168; p 315). La estimación del kernel se puede realizar en forma paramérica, asignando una función específica, o no paramétrica, mediante suavizamiento (Bivand et al., 2008, loc 2176).

El estimador de la función de intensidad toma la siguiente forma:

$$\hat{\lambda}(x) = \frac{1}{h^2 q(\|\mathbf{x}\|)} \sum_{i=1}^{n} k\left(\frac{\|x - x_i\|}{h}\right), \tag{5.2}$$

donde $\mathbf{x} \in \Re^n$ es el vector de puntos espaciales observados, $k(u)$ es una función kernel bivariada simétrica y $q(\|\mathbf{x})\|$ es una función de corrección de borde. Además, la expresión (5.2) está sujeta a la condición

$$\int_A k(u)du = 1,$$

donde A representa el área de estudio a la cual pertenecen los puntos espaciales.

La función de intensidad tiene dos usos principales: el primero es directo y permite estimar donde se presenta con mayor frecuencia cierto evento; la segunda es la posibilidad de probar hipótesis al comparar dos tipos de eventos, uno de referencia y otro de contraste.

Para el efecto de contraste de hipótesis se define la función de riesgo dada en la expresión:

$$r(x) = \log \frac{\hat{\lambda}_1(x)}{\hat{\lambda}_0(x)} \tag{5.3}$$

donde $\hat{\lambda}_0(x)$ y $\hat{\lambda}_1(x)$ son las funciones de intensidad de los eventos tipo 0 y tipo 1 respectivamente. La función de riesgo en (5.3) determina la propensión relativa a formar clusters entre los dos eventos de contraste. Como estadístico de prueba se sugiere (Waller,

2009, p 316) construir la distribución empírica de (5.3) mediante simulación y remuestreo, a partir de la cual se define la región de confianza para la hipótesis nula $\mathcal{H}_0 : \lambda_1(x) = \lambda_0(x)$.

5.3. Definición de regiones mediante *MST*

El método del *Árbol Generado Mínimo* es de los más utilizados como mecanismo para la definición de regiones contigüas. El principio básico sobre el cual se sustenta es obtener la máxima *disimilitud* entre unidades espaciales para descartar aquellas que no se consideren parte de un grupo. La configuración regional que se genere depende directamente del establecimiento de una métrica, misma que se hará acorde al contexto de estudio y la propiedades que tenga.

5.3.1. **Disimilitud**

Considere el conjunto de objetos espaciales Ω cuyos atributos son $\{A_1, A_2, \ldots, A_n\}$. Si $\mathbf{x} \in \Omega$ entonces su ávatar es $\mathbf{x}^t = (a_1, a_2, \ldots, a_n)$ donde $a_i \in A_i$. Esto es, cada objeto espacial es definido por sus características expresadas como un vector $\mathbf{x} \in \Re^n$. La topología definida en Ω genera el grafo $\mathcal{G} = (V, L)$, donde V es un conjunto de vértices (nodos) y L de aristas. La arista que conecta a los vértices v_i y v_j del conjunto V se representa como el par (v_i, v_j), la cual tendrá un costo (disimilitud) asociado $d(v_i, v_j)$.

Es importante señalar que el costo de las aristas está relacionado al contexto de definición de los objetos espaciales y la métrica utilizada. Particularmente en un contexto socioeconómico se consideran variables aleatorias continuas. En este sentido, los conceptos de similitud y disimilitud juegan un rol determinante en la construcción de clusters. En las definiciones 3 y 4 se establecen las propiedades que deben cumplir (Assuncão et al., 2006, p 800).

Definición 3 *Sea $\mathcal{E}$ un conjunto de objetos espaciales. La* similitud (proximidad) *entre los objetos r y s de $\mathcal{E}$ con avatares $\mathbf{x}_r$ y $\mathbf{x}_s$, denotada por d_{rs}, satisface las siguientes tres condiciones:*

i) $0 \leq d_{rs} \leq 1, \forall r, s \in \mathcal{E}$

ii) $d_{rs} = 1 \Leftrightarrow r \equiv s$

iii) $d_{rs} = d_{sr}$

$\boxdot$

Definición 4 *Sea $\mathcal{E}$ un conjunto de objetos espaciales. La* disimilitud *entre los objetos r y s de $\mathcal{E}$ con avatares $\mathbf{x}_r$ y $\mathbf{x}_s$, denotada por d_{rs}, satisface las siguientes tres condiciones:*

i) $d_{rs} \geq 0, \forall r, s \in \mathcal{E}$

ii) $d_{rs} = 0 \Leftrightarrow r \equiv s$

iii) $d_{rs} = d_{sr}$

$\boxdot$

La delimitación de regiones en un espacio geográfico construida a partir de una partición de este, puede ser representada mediante un grafo cuyos nodos son las unidades de referencia para aglutinar y las aristas la contigüidad. El concepto de grafo es por tanto relevante en la metodología propuesta para determinar una partición o regionalización.

Definición 5 *Sea $V = \{v_1, v_2, v_3, \ldots, v_r\}$ un conjunto de puntos en $\Re^n$ llamados vértices y $L = \{l_1, l_2, \ldots, l_s\}$ un conjunto de líneas en $\Re^n$ denominadas aristas que unen a pares de vértices. A la conjunción de V y L se le llama* Grafo *$\mathcal{G}$ y se denota como $\mathcal{G} = \{V, L\}$.*

$\boxdot$

Las medidas de disimilitud más utilizadas se definen en el apéndice A. Las aristas en un grafo estarán ponderadas por una medida de disimilitud. Entre las medidas más comunes se encuentra la norma euclideana. Dados dos los vectores $\mathbf{x}_i$ y $\mathbf{x}_j$ en $\Re^n$ la norma eculideana está dada por la expresión:

$$d_{ij} = d(v_i, v_j) = ||\mathbf{x}_i - \mathbf{x}_j|| = \sum_{k=1}^{n} (x_{i_k} - x_{j_k})^2$$

Definición 6 *Una* ruta *del nodo v_1 al nodo v_k es una sucesión de nodos $v_1, v_2, \ldots, v_k$ conectados mediante las aristas $(v_1, v_2), (v_2, v_3), \ldots, (v_{k-1}, v_k)$. Se dice que un gafo $\mathcal{G}$ es conexo si para cualquier par de nodos existe una ruta que los conecta. Un* circuito *es una ruta donde $v_1 = v_k$.*

⊡

En el apéndice A se describen otras métricas utilizadas: Euclideana ponderada, Euclideana con centroide, Mahalanobis, Manhattan, Minkowsky, Camberra y Norma suprema.

Definición 7 *Una* árbol generado *T es un grafo $\mathcal{G}$ que contiene n nodos, donde cualquier par de nodos se conecta por una sola ruta y el número de aristas es $n - 1$.*

⊡

Por sus siglas en inglés se hará referencia a un árbol generado como *ST*. Se sigue directamente de la definición 7 que a partir de la remoción de cualquiera de las aristas de un árbol generado T se obtienen dos subgrafos no conexos. Estos subgrafos son candidatos potenciales a formar un cluster, lo que equivale a una región ya que existe un isomorfismo entre una región y un árbol generado a partir de un subgrafo.

A partir de un grafo $\mathcal{G}$, donde cada una de las aristas es ponderada por una medida de disimilitud, se puede construir un árbol generado de tal forma que la suma de disimilitudes sea lo menor posible. Esto genera una propiedad deseable que servirá de base para la definición de regiones representadas por subconjuntos de nodos con un grado de afinidad preestablecido.

Definición 8 *Una* árbol generado mínimo *T de un grafo $\mathcal{G}$ es un árbol generado con costo mínimo, definido este como la suma de disimilitudes.*

⊡

Pos sus siglas en inglés se referirá a un árbol generado mínimo como *MST*.

La unicidad de un *MST* es una propiedad deseable cuando se traslada a un contexto de regionalización. En este sentido, esta característica es frecuente en la práctica donde los pesos de las aristas son todas distintas

Teorema 1 *(Unicidad del MST)*

Sea G un grafo con aristas $\{d_1, d_2, \ldots, d_n\}$. *Si se satisface que* $d_i \neq d_j, \forall i \neq j$, *entonces solo existe un MST.*

Demostración

Supóngase que existen dos MST, digamos A y B asociados al grafo G. Sin pérdida de generalidad sea a_1 *la arista de menor peso que pertenece a A pero no a B.*

Al menos debe haber una arista en C que no está en A, de otra manera no sería un ST.

Como B es un MST, entonces $\{a_1\} \cup B$ *contiene un ciclo C el cual posee una arista* a_2 *que satisface* $d_2 > d_1$. *Esto se debe a que todas las* $a_i \in B$ *con menos peso están en A mediante la inclusión de* a_1.

Reemplazando a_2 *por* a_1 *en B se generaría un árbol de menor peso, pero esto contradice el hecho de que B es un MST. Por consiguiente solo A es un MST.*

□

5.3.2. Construcción del *MST*

El algoritmo para construir un *Árbol Generado Mínimo* (*MST*) asociado al espacio económico es relativamente sencillo, el menos en su definición; este se basa en determinar a partir de un nodo dado las disimilitudes de sus vecinos, entendiendo como vecinos a los objetos espaciales contigüos o a su equivalente que son las unidades económicas colindantes.

Dos de los algoritmos principales son el de *gráfico transversal de primer plano* y *gráfico transversal de primera extensión* (Mihalcea y Radev, 2011, loc 525-579). El más común es el primero y será el que se aplicará en este trabajo.

Sea $V = \{v_1, v_2, \ldots, v_n\}$ un conjunto de nodos, $l = (v_i, v_j)$ la arista de v_i a v_j, T_k un árbol *MST* con k nodos y $d(v_i, v_j) = ||v_i - v_j||$ la distancia entre nodos. El algoritmo transversal en primer plano para construir el *MST* asociado a V es como sigue:

1. seleccionar cualquier $v_i \in V$ y definir $T_k = T_1 = (\{v_i\}, \emptyset)$;
2. encontrar $l' = (v_i, v_j)$, $\forall v_j \notin T_k$ de tal manera que $d(v_i, v_j) \leq d(v_i, v_k)$, $\forall v_k \notin T_k$;
3. definir $T_{k+1} = T_k \cup \{v_j, l'\}$;
4. repetir el segundo paso hasta tener $T_k = T_n$.

En al figura 5.1 se muestra el diagrama de flujo para la construcción del *MST*. Partiendo del supuesto de la continuidad en las variables que definen las características de los objetos espaciales, es difícil que una vez seleccionado un nodo inicial se encuentre más de un nodo con distancia mínima, posibilidad que se reduce aún más con el incremento de la dimensión del vector de características. Más aún, esto cambiará también en función de la métrica que se utilice, en este caso se asume la distancia euclidiana como medida de disimilitud.

5.3.3. Partición del *MST*

Una vez construido el *MST* y entendido como un isomorfismo de la estructura socio-económica a regionalizar, el siguiente paso es aprovechar las propiedad de esta estructura abstracta para generar una partición que será equiparable a la obtención de regiones en el contexto de estudio.

Teorema 2 *Dadas las características del MST, la remoción de una de las aristas deriva en dos sub-grafos MST conexos sin circuitos, esto es, se tendrán dos árboles no conectados entre si.*

□

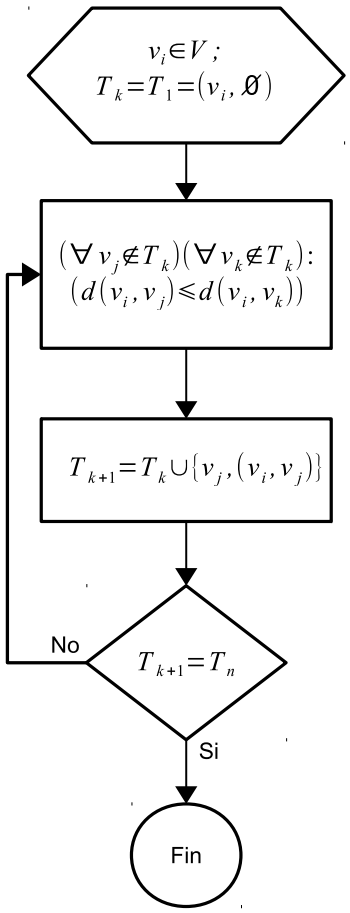


Figura 5.1: *Algoritmo para generar un MST*

Cada uno de los sub-grafos corresponderá en su contraparte del espacio económico a una región. Por tanto, para lograr k sub-árboles (regiones) se tendrían que eliminar $k-1$ aristas.

Para obtener la segmentación del *MST* se deben establecer condiciones generales que sirvan como punto de partida. El problema central en el establecimiento de una partición radica en el criterio que se deba aplicar para considerar a un árbol como un conjunto homogéneo en relación a sus nodos y el contexto de lo que representan las aristas expresado como una métrica.

Un criterio de optimalidad para la partición del *MST* se basa en minimizar la suma de las desviaciones al cuadrado intracluster, esto es, obtener el mínimo de la siguiente función:

$$Q(\mathcal{P}) = \sum_{i=0}^{k} SCD_i, \tag{5.4}$$

donde $\mathcal{P}$ representa la partición del *MST* en k árboles, $Q(\mathcal{P})$ es la medida de calidad de la partición y SCD_i es la suma del cuadrado de las desviaciones en la región i. Esta suma se expresa como sigue:

$$SCD_k = \sum_{j=1}^{m} \sum_{i=1}^{n_k} (x_{ij} - \bar{x}_j)^2, \tag{5.5}$$

donde n_k es la cantidad de objetos espaciales en *k-ésimo* árbol y x_{ij} es el *j-ésimo* atributo del objeto espacial i.

La primera parte será expresar este procedimiento en notación matricial. Para tal efecto, sea $\mathbf{W} \in \Re^{a \times n}$ la matriz cuyas columnas son los vectores de características de los n nodos (objetos espaciales):

$$\mathbf{W} = [\mathbf{w}_1 | \mathbf{w}_2 | \cdots | \mathbf{w}_n], \quad \mathbf{w}_i \in \Re^a, i = 1, 2, \ldots, n$$

Para una partición dada, bajo la restricción de contigüidad, suponga que se tienen c clusters expresados como bloques de $\mathbf{W}$:

$$\mathbf{W} = [\mathbf{W}_1 | \mathbf{W}_2 | \cdots | \mathbf{W}_c], \quad \mathbf{W}_i \in \Re^{a \times n_k}, i = 1, 2, \ldots, c$$

Sin pérdida de generalidad se considera el k-ésimo cluster:

$$\mathbf{W}_k = [\mathbf{w}_{k_1} | \mathbf{w}_{k_2} | \cdots | \mathbf{w}_{k_{n_k}}], \quad \mathbf{w}_{k_i} \in \Re^a, i = 1, 2, \ldots, n_k$$

Dentro del cluster k se calcula el promedio de cada una de las características de todos los nodos, dando como resultado el vector $\bar{\mathbf{w}}_k$:

$$\bar{\mathbf{w}}_k = \begin{pmatrix} \bar{x}_{k_1} \\ \bar{x}_{k_2} \\ \vdots \\ \bar{x}_{k_a} \end{pmatrix}$$

donde $\bar{x}_{k_i}$ representa el promedio de la característica i de todos los nodos del cluster k. Además, sea $\mathbf{M}_k$ la matriz:

$$\mathbf{M}_k = \mathbf{1} \otimes \bar{\mathbf{w}}_k, \quad \mathbf{1} = \underbrace{(1, 1, \ldots, 1)}_{n_k}$$

donde $\otimes$ es el producto *kronecker*. Utilizando las matrices $\mathbf{W}_k \in \Re^{a \times n_k}$ y $\mathbf{M}_k \in \Re^{a \times n_k}$ se obtiene la matriz centrada en las características del cluster k:

$$\mathbf{W}_{k_0} = \mathbf{W}_k - \mathbf{M}_k$$

la cual concentra las desviaciones de los nodos respecto a al promedio de cada la característica correspondiente. A partir de esta matriz se define $SD_k = tr(\mathbf{W}_{k_0}\mathbf{W}_{k_0}^t) = \sum_{i=1}^{n_k} ||\mathbf{w}_{k_i}||^2$.

Finalmente, la medida de la calidad de la partición (regionalización) se obtiene a partir de la expresión (5.6).

$$Q(\mathcal{P}) = \sum_{i=1}^{c} SD_i = \sum_{i=1}^{c} tr(\mathbf{W}_{i_0}\mathbf{W}_{i_0}^t) \tag{5.6}$$

5.4. Termodinámica

La entropía en el contexto de la termodinámica establece una estrecha conexión con el concepto de información planteado por Shannon. En este sentido, la relevancia del concepto de entropía queda plasmado en la segunda ley de la termodinámica que establece lo siguiente (Bryant, 2012, p. 59):

No es posible construir un sistema que opere en un ciclo, extraiga calor de su contenedor y genere una cantidad equivalente de trabajo en el entorno.

Lo que esencialmente se plantea aquí es que existe energía libre que no es utilizada cuando se realiza un trabajo. Esto significa que en un sistema no se podrá tener una eficiencia del 100 % en el uso de la energía disponible.

5.4.1. **Entropía**

En termodinámica la entropía puede ser considerada como una medida de la energía que no puede ser utilizada para generar trabajo. Su contraparte en mecánica estadística la define en términos de la probabilidad que que un sistema esté en un estado, la cual es referida como la *aleatoreidad*, en el sentido de distribución uniforme, expresada en un sistema (Bryant, 2012, p 61), también conocida como *entropía de Boltzmann*.

En referencia a un sistema cerrado de n partículas en el cual se distinguen k estados posibles donde cada una de estas partículas puede residir, se plantea la definición 9 (Georgescu-Roegen, 1999, p 144).

Definición 9 *La entropía de Boltzmann está dada por la expresión*

$$S = k \ln W \tag{5.7}$$

donde $k = 1.38 \times 10^{-16}$ *es la constante de Boltzmann y*

$$W = \frac{n!}{n_1! n_2! \cdots n_k!} \tag{5.8}$$

□

Notemos que W es la cantidad de posibilidades en las que puede aparecer cada configuración (distribución) de las n partículas en los k estados.

Aplicando logaritmo en ambos lados de la expresión (5.8) y utilizando la aproximación de *Stirling* para factoriales se puede demostrar que

$$\ln W = -\sum_{i=1}^{k} n_i \ln\left(\frac{n_i}{n}\right)$$

Entonces, haciendo $f_i = \frac{n_i}{n}$ se sigue que $S = -kNH$, donde

$$H = \sum_{i=1}^{k} f_i \ln f_i \tag{5.9}$$

La función H en (5.9) es utilizada por *Boltzmann* en el enfoque estadístico de la termodinámica. Además, $\frac{S}{n} = -kH$ se interpreta como la entropía promedio en el sentido de *Boltzmann* y, más aún, $\frac{S}{n}$ también es proporcional a la entropía en el sentido de *Shannon* en un enfoque de información. Esto se puede ver si denotamos $\mathcal{H} = -H$, de donde se obtiene $\frac{S}{n} = k\mathcal{H}$.

5.4.2. **Termoeconomía**

La segunda ley de la termodinámica (acotación en la sección 5.4) implica una analogía asociada al crecimiento económico, la cual en un contexto de la denominada termoeconomía, establece que

> *La conversión de energía y la producción de entropía determina el crecimiento del bienestar.*

Esta relación no solo se restringe a una relación general sino que es llevada a un nivel de analogía de varias maneras. Ya el economista Paul Samuelson desde 1947 comparaba la presión y volumen en física con el precio y volumen en economía, haciendo uso del principio de Chatelier que, como lo describe Bryant (2012, p 4), establece que

> *"Si un cambio ocurre en alguno de los factores en los cuales se basa el equilibrio de un sistema, el sistema tenderá a autoajustarse para anular, en la medida de lo posible, los efectos de este cambio"*.

La principal analogía que se hace de la economía con la termodinámica se relaciona con el comportamiento de un gas ideal, donde se establece la relación

$$PV = NkT \tag{5.10}$$

donde P es el volumen del gas, P la presión ejercida sobre las paredes de un contenedor, N la cantidad de moléculas, k la constante de *Boltzmann* y la temperatura T entendida como una medida de energía cinética. Esta relación puede circunscribirse a dos entornos, a saber, un sistema al considerarse cerrado o uno con flujo al existir intercambio con otro entorno.

La contraparte en economía de la expresión (5.10) requiere de establecer la equivalencia de estos parámetros. La constante k es entendida como el contenido productivo que cada unidad económica posee; se considera de esta manera ya que en algunos casos se interpreta como el valor monetario, el cual no necesariamente es única o estandarizada en diversas economías. A diferencia de su contraparte física, en economía el valor de k no es constante y depende del producto de referencia.

En economía se puede considerar que existen las dos maneras de concebir a un gas: como contenido en un repositorio y como flujo en unidades que transitan de un repositorio a otro. En este sentido, se considera a P como el precio, V_t al volumen de flujo en cierto periodo de tiempo y a T_t como el índice de valor de mercado, dando como resultado la relación:

$$PV_t = NkT_t \tag{5.11}$$

5.5. Información

5.5.1. **Entropía de Shannon**

Cover y Thomas (2006, p 14) define la entropía de Shannon como una medida de incertidumbre. Técnicamente esta dada como sigue:

Definición 10 *Sea X una variable aleatoria discreta con dominio en $\mathcal{X}$ y función de densidad de probabilidad $f_x(x) = P[X = x] = p(x)$, para $x \in \mathcal{X}$. La* ***Entropía*** *$H(X)$ se define como*

$$H(X) = -\sum_x p(x) \log_2[p(x)] \tag{5.12}$$

donde se define $(0)\log_2(0) = 0$.

Una interpretación a esta medida es la cantidad de incertidumbre que presenta una variable aleatoria discreta. Por ejemplo, supongamos que un experimento u observación de un fenómeno puede derivar en dos resultados; estos resultados se codifican mediante la variable aleatoria X con valores posibles de $X = 0$ o $X = 1$. Además, si la probabilidad de que X tomo el valor de 1 es p, esto es, $P\{X = 1\} = 0$, entonces X sigue una distribución *Bernoulli* con parámetro p, denotado por $X \sim Ber(p)$, cuya función de densidad está dada por la siguiente expresión:

$$f_x(x) = p^x(1-p)^{1-x} I_{\{0,1\}}(x)$$

Si en particular $X \sim Ber(0.5)$, entonces, evaluando en (5.12), se tiene $H(0.5) = -(0.5)\log_2(0.5) - (1-0.5)\log_2(1-0.5) = 1$ y $H(1) = -(1)\log_2(1) - (1-1)\log_2(1-1) = 0$. Esto es, cuando $p = 1$ entonces hay un solo resultado factible y por lo tanto se tiene una incertidumbre 0, mientras que cuando $p = 0.5$ es igualmente probable que ocurra cualquiera de los resultados por lo que se tiene máxima incertidumbre, en este caso con valor 1.

Otra interpretación más reciente de la entropía de *Shannon* es considerarla como una medida cuantitativa de *distiguibilidad* de un subsistema en relación al sistema en el que se encuentra inmerso (Ayres, 1994, p 44). Este enfoque tomará relevancia en el contexto de regionalización de máxima información.

5.5.2. Divergencia de Kullback-Leibler

Si bien existen pruebas que permiten probar la hipótesis de igualdad entre dos distribuciones de probabilidad, estas no establecen una medida de proximidad entre dichas distribuciones. Este problema usualmente se enfoca a probar si la distribución de probabilidad de cierta variable aleatoria de interés se considera estadísticamente igual a una distribución teórica esperada. Se rechace o no la hipótesis de igualdad es de interés establecer la proximidad entre una distribución observada y una hipotética.

Supongamos que una muestra aleatoria X proviene de una población cuya distribución de probabilidad es $G_x(x)$, la cual se quiere contrastar contra una distribución arbitraria $F_x(x)$. Asumiendo que ambas distribuciones son discretas con densidades $g_x(x)$ y $f_x(x)$ respectivamente, se considera que la bondad del modelo $f_x(x)$ es determinada por la proximidad a la distribución verdadera $g_x(x)$.

La *entropía relativa* (Cover y Thomas, 2006, p 19), derivada de la definición 10 permite establecer la distancia entre distribuciones de probabilidad. Este concepto juega un rol central para establecer la similitud o proximidad entre regiones económicas.

Definición 11 *La **entropía relativa** o **distancia de Kullback-Leibler** entre dos funciones de masa de probabilidad $g_x(x)$ y $f_x(x)$ se define como*

$$\begin{aligned} D(g||f) &= \sum_{x\in\mathcal{X}} g_x(x)\log_2\frac{g_x(x)}{f_x(x)} \\ &= E_g \log_2\frac{g_x(x)}{f_x(x)} \end{aligned} \tag{5.13}$$

donde se define $(0)\log_2\frac{0}{0} = (0)\log_2\frac{0}{f_x} = 0$ *y* $(g_x)\log_2\frac{g_x}{0} = \infty$.

⊡

A la entropía relativa también se le denomina información de Kullback-Leibler o divergencia de Kullback-Leibler (Konishi y Kitagawa, 2008, p 29). Aunque esta medida,

también llamada *distancia de Kullback-Leibler (DKL)*, no es una métrica auténtica si aporta una medida unidireccional; bajo ciertas circunstancias se puede calcular el promedio de las distancias unidireccionales entre dos distribuciones de probabilidad y ser utilizada como métrica.

Una extensión a variables aleatorias continuas es dada por (Konishi y Kitagawa, 2008, p 23).

Definición 12 *La **información de Kullback-Leibler** entre dos funciones de masa de probabilidad $g_x(x)$ y $f_x(x)$ se define como*

$$I(g||f) = E_G\left[log_2\frac{g_x(x)}{f_x(x)}\right] = \begin{cases} \sum_{x\in\mathcal{X}} g_x(x)\log_2\frac{g_x(x)}{f_x(x)} & \textit{para un modelo discreto} \\ \int_{-\infty}^{\infty} g_x(x)\log_2\frac{g_x(x)}{f_x(x)} & \textit{para un modelo continuo} \end{cases} \tag{5.14}$$

A partir de la definición 12 se puede demostrar (teorema 8, apéndice B.2) que Si $F_X \sim U(\mathcal{X})$ entonces $I(g||f) = log_2|\mathcal{X}| - H_g(X)$.

El teorema 8 determina una medida de proximidad de una distribución G_x a una distribución uniforme U_x definida en el dominio $\mathcal{X}$. En el contexto de distribución de ingresos, la relevancia de este resultado radica en que se puede interpretar como una medida del grado de desigualdad, esto debido a que si $G_x = U_x$ entonces la información $I(g||u) = 0$, lo que significa que el ingreso estará igualmente distribuido en las categorías que se definan.

Para este trabajo la división de la medida $\mathcal{X}$ se hará en ocho partes iguales a lo largo de su rango. La razón de seleccionar esta cantidad, conjuntamente con la base 2 del logaritmo, radica en que se tendrán dos propiedades deseables:

1. el valor máximo de información es el entero 3;

2. la unidad de información es el bit.

La primera de las propiedades da en forma implícita una manera directa de escalar al rango $(0,1)$. La segunda propiedad establece una medida de información ampliamente utilizada en otros contextos, particularmente en el área de computación.

Se observó que la entropia media en el sentido de *Boltzmann*, cuando se considera la distribución de partículas en s estados posibles, tiene una estrecha relación con la entropía en un enfoque de teoría de información, específicamente esta es el negativo de la entropía media en términos de la función propuesta por *Shanon* (Cover y Thomas, 2006, p 14).

Otra manera de plantear la entropía de *Boltzmman* es considerar la distribución de frecuencias relativas, esto es, analizar la distribución de probabilidad de las partículas cuando se asume que cada estado tiene cierta probabilidad de ser ocupado. Para esto, consideremos un gas como un sistema de n partículas que interactúan de tal manera que cada una de estas partículas se ubica en uno de s estados posibles. Si $\{n_1, n_2, \ldots, n_s\}$ es la distribución de frecuencias verdadera de partículas en lo s estados y $\{g_1, g_2, \ldots, g_s\}$ su respectiva distribución de frecuencias relativas, donde $g_i = \dfrac{n_i}{n}$, entonces la probabilidad de que se dé esta configuración, en relación a una distribución arbitraria $\{f_1, f_2, \ldots, f_s\}$, es

$$W = \frac{n!}{n_1! n_2! \cdots n_k!} f_1^{n_1} f_2^{n_2} \cdots f_k^{n_k} \tag{5.15}$$

Tomando logaritmos a ambos lados de la expresión (5.16) y aplicando la aproximación de *Stirling* para factorial, se puede demostrar (ver apéndice B.1) que la entropía es una cantidad que varía proporcionalmente a la probabilidad W en (5.16); específicamente se tiene que

$$\log W \approx -n \sum_{i=1}^{s} g_i \log\left(\frac{g_i}{f_i}\right) \tag{5.16}$$

Se sigue entonces que es proporcional a la información de *Kullback-Leibler*, esto es, $B(g,f) \propto I_{kl}(g,f)$. Más aún, la relación entre ambos enfoques es de signo contrario lo que significa que a mayor entropía en el sentido de *Boltzmann* menor en el sentido *Kullback-Leibler*.

Si f sigue una distribución uniforme en $(0,1)$, denotado como $f \sim U(0,1)$, entonces se

tendrá una medida de proximidad (o disimilitud) de g a un comportamiento uniforme. Este resultado toma relevancia en un contexto económico si consideramos la distribución del ingreso en cierta unidad económica para establecer su proximidad a una distribución equitativa (equiprobable) del mismo (Georgescu-Roegen, 1999, p 6).

5.6. Construcción del *MST* como función de la divergencia de Kullback-Leibler

En un problema regionalización se debe utilizar una medida consistente con el espacio geográfico, en el cual se introduce también una medida relativa al fenómeno de estudio. En términos de dimensión y agregación, deberá ser dimensionalmente adecuada además de poderse desagregar.

Se hará énfasis en una medida que permita establecer el grado de disimilitud de un objeto espacial en relación a objetos colindantes geográficamente y, en un contexto de sistemas, establecer en qué dimensión un objeto espacial en particular es distinguible de su entorno, entendido este como el conjunto espacial en el cual se encuentra inmerso.

Las características de una medida que garantizan las propiedades deseables mencionadas, aparecen en forma intrínseca en la medida original de información de Shannon, la cual se define en términos de la distribución de probabilidad del fenómeno de estudio (Batty, 1978, p 118):

- debe ser una función continua de las probabilidades asociadas a la distribución;
- debe ser monótona creciente en la cantidad de eventos, intervalos o zonas;
- debe tener la propiedad de aditividad.

La primera de las propiedades garantiza que se considere a la estructura probabilística de las características de estudio en lugar de basarse en valores agregados.

La segunda propiedad se asocia a la resolución de la regionalización ya que dependiendo de esta se tendrá más información ante una desagregación más fina en términos de la

cantidad de subconjuntos considerados como regiones.

La tercera propiedad hace posible la agregación lineal en el espacio de probabilidad de la distribución del fenómeno de referencia. Este comportamiento de la medida elimina residuos asociados a información no explicada en el contexto de la regionalización.

La construcción de un conjunto de una partición de nodos isomorfa al problema de regionalización se lleva al cabo en dos etapas: la primera es la construcción del *MST* que conecta a todos los nodos; la segunda se refiere a la remoción de aristas para generar *sub-MST* al grado de resolución que se establezca medida como nivel de desagregación.

La primera parte se obtiene mediante el desarrollo del algoritmo basado en el concepto de *MST* (Assuncão et al., 2006, p 800) mediante el uso de métricas más comunes (apéndice A). Para tal efecto se realizan las siguientes etapas:

1. Se simulan dos poblaciones basadas en dos características las cuales siguen distribuciones normales;
2. se define la distancia (disimilitud) a utilizar;
3. como criterio heurístico se seleccionan las dos unidades espaciales que tengan la mínima disimilitud los cuales son etiquetados con un índice y forman parte del conjunto base para el *MST* y son unidos por una arista. A diferencia de Assuncão et al. (2006, p 801) que selecciona un primer vértice en forma arbitraria, aquí se selecciona el de mínima disimilitud entre toda el conjunto de vértices.;
4. se calculan las disimilitudes de estos dos nodos con relación al resto de los nodos libres y se selecciona aque con menor disimilitud, se integra al conjunto de índices y se une con una arista;
5. se repite el proceso hasta agotar todos los nodos.

Una vez que se tiene el *MST* base, para la segunda etapa de construcción de sub-*MST* se establece un criterio de remoción de aristas. Dato que no se tienen circuitos[2], es claro

[2]Si se traza una ruta desde un nodo a través de aristas, no se puede retornar a ninguno de los nodos sin recorrer más de una vez alguna arista.

que por cada eliminación de una arista se obtienen dos *MST* a los que se denomina sub-árbol mínimo. Inicialmente se puede considerar que todo el conjunto de unidades espaciales conforman una región; al remover una arista se crean dos regiones, repitiendo este procedimiento por cada remoción se genera una nueva regionalización con más resolución.

El criterio general para remover una arista es de tal manera que la división generada derive en un incremento en la calidad total de los clusters resultantes. En esta caso la calidad se calcula como la suma de las desviaciones al cuadrado intracluster dada en las expresiones 5.4 o 5.6, criterio referido también como *intramax* (Masser y Brown, 1978b, p 156) . Posteriormente esta medida de calidad se extenderá al concepto de máxima información.

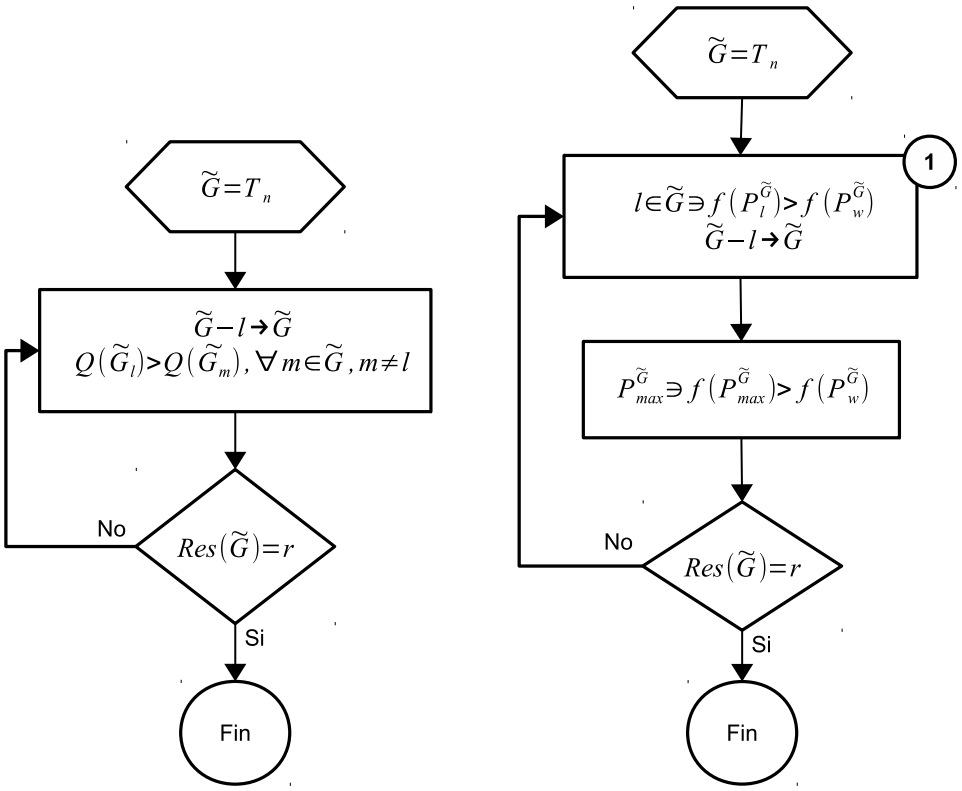
EL diagrama de flujo en la figura 5.2(a) muestra la metodología a seguir para determinar una partición del *MST* de máxima calidad en el sentido de la norma de referencia.

Los pasos a seguir son:

1. Tomar todo el *MST* (T_n) como una región a la que denominamos $\tilde{G}$, esto es $T_n \rightarrow \tilde{G}$;
2. seleccionar una arista $l \in \tilde{G}$ a remover que cumpla con la condición de ser la que mayor calidad genere comparada con cualquier otra arista $m \in \tilde{G}$;
3. actualizar $\tilde{G}$ a la partición generada: $\tilde{G} - l \rightarrow \tilde{G}$;
4. repetir el proceso hasta llegar a una resolución (desagregación) previamente establecida.

Aunque este principio de división jerárquica en que se fundamenta el algoritmo de partición de un *MST* es básico no está exento de de cierta complejidad en su aplicación. Si se aplicara de forma exhaustiva el algoritmo podría ser costoso en términos de tempo de cómputo.

El principio aplicado se puede enfocar como la obtención de un óptimo dado una función objetivo (figura 5.2(b)). Si la remoción de la arista l del árbol T deriva en dos árboles T_1 y T_2 la función objetivo se especifica en la ecuación(5.17).



(a) ***Remoción de aristas en un MST*** **(b)** ***Remoción con función objetivo***

Figura 5.2: *Remoción de aristas en un árbol generado mínimo*

$$f_1(\mathcal{P}_l) = SCD_T - (SCD_1 + SCD_2) \tag{5.17}$$

donde $\mathcal{P}_l$ se refiere a la partición generada al remover la arista l.

Dado que la búsqueda de aristas a remover en forma exhaustiva es un problema *NP-complejo* este puede transformarse para hacer eficiente el proceso. Para tal efecto, Assuncão et al. (2006, p 804) presenta un criterio heurístico basado en el análisis de candidatos vecinos ya visitados.

El procedimiento consiste en buscar en el espacio de soluciones $G = \{G_1, G_2, \ldots, G_{n-1}\}$ el mejor candidato, asumiendo en este caso que es G_l obtenido a partir de la remoción de la arista l. Si $G_l = \{G_{l_1}, G_{l_2}, \ldots, G_{l_k}\}$ el conjunto de aristas generadas por los vértices vecinos a los de la arista l, entonces se evaluará (5.17) en estas opciones para seleccionar

aquella que la maximice.

La mejor solución en cada paso puede corresponder solo a un óptimo local, siendo necesario extender el criterio de decisión para garantizar al búsqueda de óptimos globales. Para tal efecto, se introduce una segunda función objetivo que homogeniza los árboles obtenidos en la división de tal manera que estén más balanceados en términos de su dimensión. Tal función es de la forma

$$f_2(\mathcal{P}_l) = min[(SCD_T - SCD_{l_a}), +(SCD_T - SCD_{l_b})] \tag{5.18}$$

donde l_a y l_b son dos vecinos de l, esto es, $l_a, l_b \in G_l$.

La evolución en la búsqueda de soluciones depende de la primera elección de arista a remover. Un criterio a aplicar es la ubicación del centro de G, el cual es considerado como aquel que divide en dos árboles de similar tamaño. Si bien este criterio es razonable intuitivamente, es fundamental considerar que el centro en este caso estaría condicionado por el peso de las aristas, las cuales no necesariamente presentarían la misma distribución tomando como único criterio la similitud de tamaños. Se aplica otro criterio heurístico, el de no remover hojas de árboles. Y en una segunda fase, aquellas regiones de dos unidades pueden integrarse a la región menos divergente.

El algoritmo general plenteado aquí se denomina Análisis de Clusters Espaciales mediante Remoción de Aristas de Árbol y es referido como *SKATER* por sus siglas en inglés. Esta método se divide en dos partes:

1. Obtener el árbol generado mínimo *MST* de adyacencia basado en las disimilitudes entre atributos de las unidades espaciales;

2. generar particiones para definir clusters espaciales usando algún criterio global de eliminación de aristas;

3. dependiendo de la evolución en el proceso de partición y tomando en cuenta el contexto de estudio es factible considerar un criterio local de eliminación de artista.

La primera parte se divide en dos condiciones: una es la la contigüidad geográfica y la otra una medida de disimilitud. Mientras que la primera es trivial en su establecimiento, la segunda es crucial para la definición de regiones.

Tomando como referencia la definición 4 de disimilitud y la diveregencia de Kullback-Leibler dada en la definición 11, notamos que esta cumple las primeras dos condiciones requeridas pero la tercera asociada a simetría no se satisface. Sin embargo, esta deficiencia se solventa si calculamos la distancia promedio entre los vértices v_i y v_j en en ambos sentidos, esto es, se define la medida de disimilitud siguiente:

$$d(v_i, v_j) = \frac{1}{2}[d_{KL}(v_i, v_j) + d_{KL}(v_j, v_i)] \tag{5.19}$$

En la segunda parte se utiliza remoción basado en máxima divergencia, contrario a la primera etapa que se utiliza la mínima disimilitud o divergencia para incorporar una arista al *MST*. Una condición adicional de tipo heurístico que se incorpora es la de no remover aristas que aíslen a un solo nodo, lo que implica que cada estrato tendrá finalmente al menos dos unidades espaciales de estudio

La prueba de eficiencia del algoritmo se hará mediante simulación.

Entre los algoritmos heurísticos más exitosos se encuentran los de *partición de grafos multinivel*, los cuales en forma recursiva generan grafos más pequeños que preservan la estructura inicial. El principio se centra en que, una vez hecha una partición, se hará una búsqueda local para repetir el procedimiento hasta lograr cierto nivel de eficiencia o bien de resolución.

Los algoritmos *KaHIP*[3] son una familia de algoritmos de partición de árboles. La versión *KaFFPa* reduce el árbol y aplica búsquedas locales; la variante *KaFFPaE* es de tipo evolutivo y mejora algunos aspectos en el criterio de búsqueda local; el procedimiento *KaBaPE* optimiza la búsqueda local, realizando un balanceo, donde para ciertas condiciones de la estructura del árbol es mejor que los otros dos métodos.

[3]Kalrushe High Quality Partitioning.

5.6.1. Simulación de regiones en Coahuila (caso unidimensional)

Como ejercicio previo para la aplicación del método de regionalización, se realizan simulaciones de la distribución de probabilidad de una característica aplicada al estado de Coahuila.

Para tal efecto, se requieren de varias etapas para, en primera instancia, hacer una clasificación de municipios en función de la distribución de probabilidad empírica de una variable simulada:

1. Se construye la matriz de contigüidad municipal para el estado de Coahuila, lo cual se detalla en la sección 6.2;
2. se considera al municipio como un conjunto de localidades, generando 38 cantidades uniformes entre 75 y 350 las cuales se asignan a cada uno de los municipios;
3. se genera un vector de 1000 datos con distribución normal estándar, la cual es dividida en quintiles;
4. tomando como referencia las 5 regiones administrativas de Coahuila se clasifican a los municipios, asignando a cada uno de estos una muestra aleatoria de los datos simulados de cada uno de los quintiles. Esto garantiza que tienen distinta distribución empírica por región (figura 5.3);
5. se obtiene la distancia media (bidireccional) de *Kullback-Leibler* de cada municipio con respecto a la entidad;
6. se obtiene la distancia de *Kullback-Leibler* de cada municipio respecto a una distribución uniforme.

La distancia *KL* estimada entre regiones se resume en el cuadro 5.1.

Se observa en las figuras 5.4 y 5.5 que utilizando la distancia *KL* preserva más la continuidad inducida en la simulación respecto a una distribución uniforme que la observada en relación a la entidad. Esta situación deberá corroborarse en la aplicación utilizando la eficiencia, basada en la entropía, para cada regionalización.

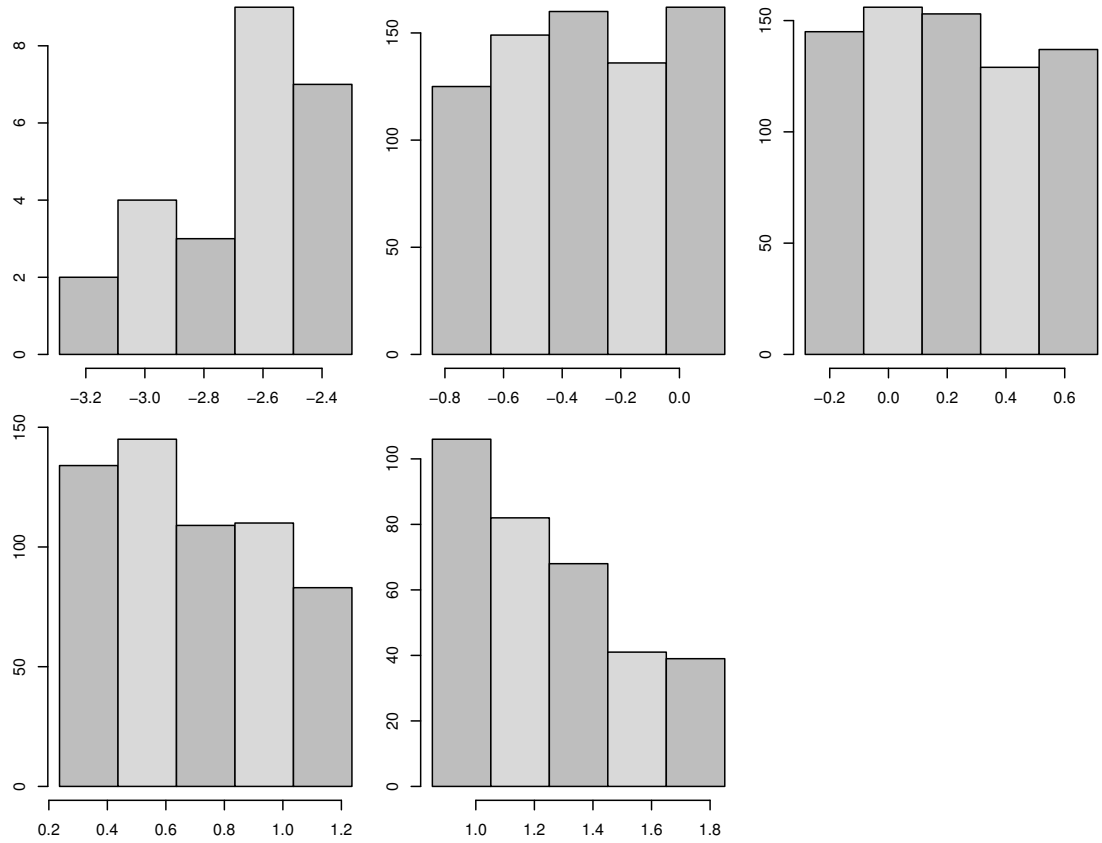


Figura 5.3: *Simulación de distribución en regiones de Coahuila*

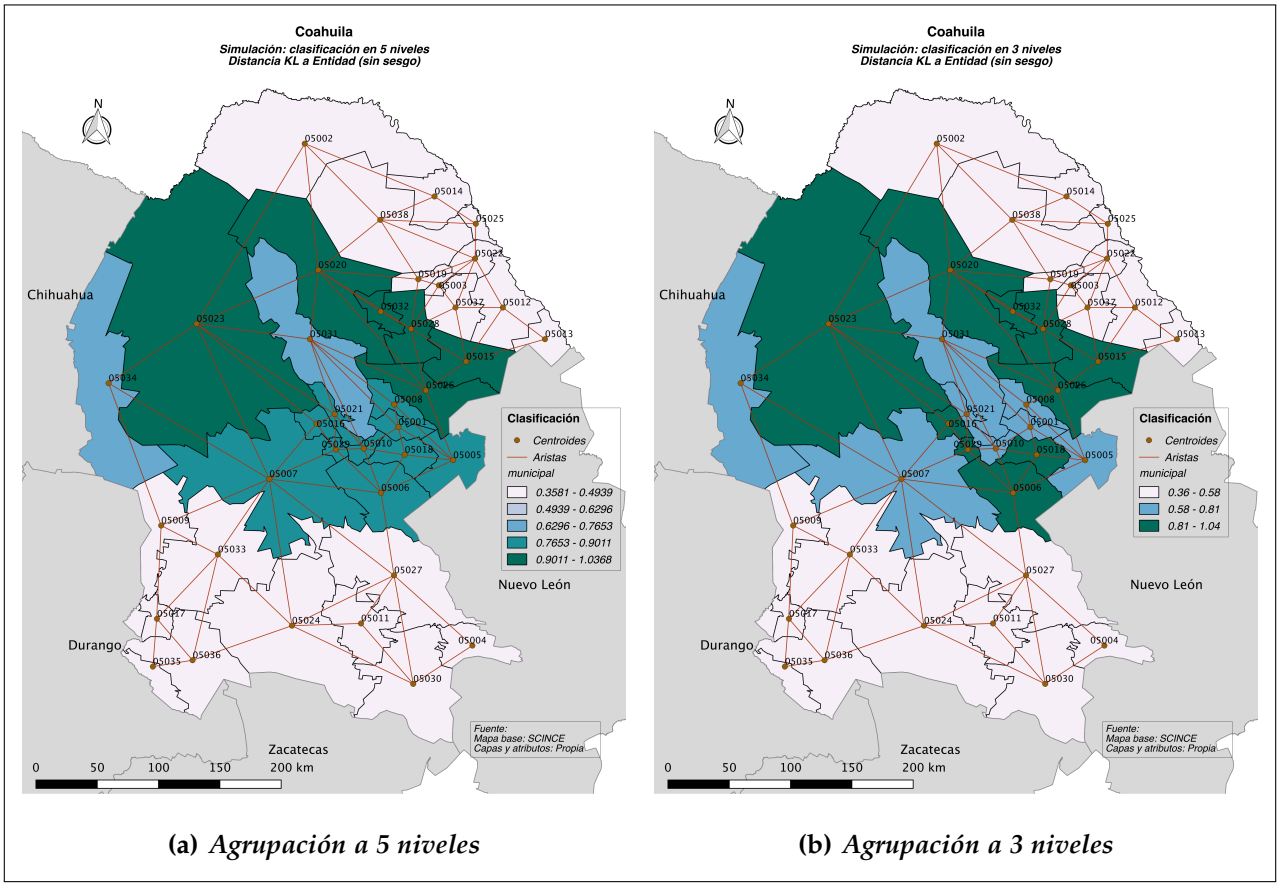
5.7. Una medida de divergencia para la construcción del *MST*

Tomando como referencia la distancia propuesta en la ecuación (5.19) que utiliza a la distribución uniforme como punto espacial común, se debe de tomar en cuenta la asimetría de las distribuciones ya que no se tendría una relación de transitividad de las distancias entre las distribuciones empíricas de referencia. Esto es, si U representa a la distribución uniforme, X_1 la primera unidad de referencia y X_2 la segunda, entonces las distancias $d(X_1, U)$ y $d(X_2, U)$ no aportan información directamente acerca de la distancia $d(X_1, X_2)$ debido a que no se considera la asimetría de las distribuciones.

El problema planeado se representa en la figura 5.6. Los triángulos representan la forma de las distribuciones de las unidades espaciales, mientras que el rectángulo es la distribución uniforme de referencia.

Cuadro 5.1: *Distancia K-L en datos simulados*

NR	KLE	KLU
1	0.653	0.109
2	0.466	0.007
3	0.398	0.007
4	0.426	0.012
5	0.520	0.067

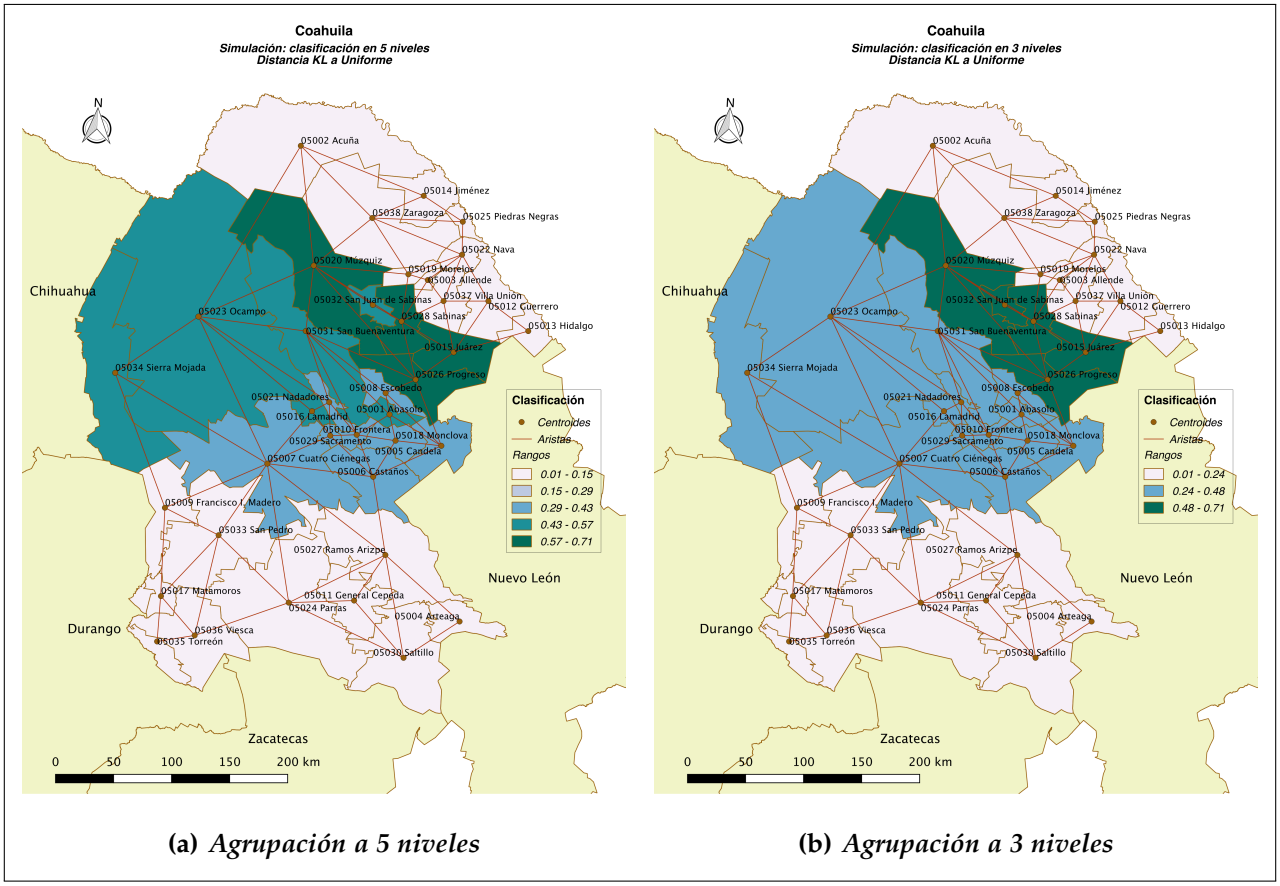


(a) ***Agrupación a 5 niveles*** **(b)** ***Agrupación a 3 niveles***

Figura 5.4: *Clasificación de municipios en simulación (Entidad)*

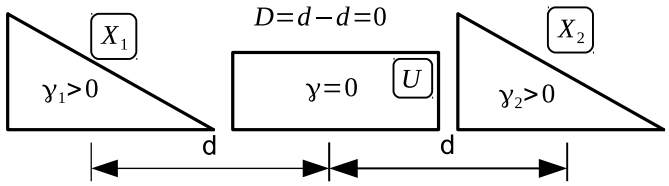
Cuando se tiene el mismo signo en la asimetría y coinciden las distancias respecto a la distribución uniforme, entonces la distancia es cero (figura 5.6(a)), pero si el signo es contrario entonces la distancia se duplica (figura 5.6(b)).

Para solventar esta incidencia en el cálculo de la distancia entre dos distribuciones (X_1

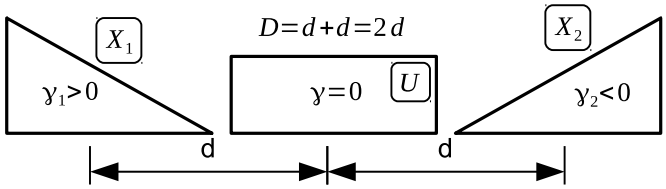


(a) ***Agrupación a 5 niveles*** **(b)** ***Agrupación a 3 niveles***

Figura 5.5: *Clasificación de municipios en simulación (Uniforme)*



(a) ***Igual signo de la asimetría***



(b) ***Distinto signo de la asimetría***

Figura 5.6: *Asimetría en distribuciones empíricas*

y X_2) se propone la divergencia dada en la definición 13 como alternativa de disimilitud:

Definición 13 *(Divergencia ϕ).*
La distancia $\phi(X_1, X_2)$ entre dos unidades espaciales expresadas como la distribución empírica de una característica de interés está dada por la siguiente expresión:

$$\phi(X_1, X_2) = |sgn(\gamma_1)div(X_1, U) - sgn(\gamma_2)div(X_2, U)|, \tag{5.20}$$

donde γ representa la asimetría, $sgn(\gamma) = -1$ para asimetría negativa, $sgn(\gamma) = 1$ para asimetría positiva, U una distribución uniforme y

$$div(X, U) = \sum_{i=1}^{k} \frac{1}{k} \log \frac{1}{kf(x_i)}$$

A esta distancia se le referirá como divergencia ϕ.

La métrica propuesta en (5.20) tiene las siguientes propiedades:

1. $\phi(X_1, X_2) \geq 0$;
2. $\phi(X_1, X_2) = 0$ si $\hat{F}_{X_1} = \hat{F}_{X_2}$, donde $\hat{F}_{X_1}$ y $\hat{F}_{X_2}$ son las distribuciones empíricas de X_1 y X_2 respectivamente;
3. es simétrica, esto es, $\phi(X_1, X_2) = \phi(X_2, X_1)$.

Con estas propiedades, (5.20) establece una media más *fina* entre dos estratos comparada con la que aportaría una medida agregada. La metodología para su cálculo es:

1. Obtener la asimetría de cada una de las distribuciones empíricas las que se denominarán γ_1 y γ_2;
2. calcular las divergencias respecto a U de las distribuciones empíricas X_1 y X_2, denotadas respectivamente como $div(X_1, U)$ y $div(X_2, U)$;
3. calcular la distancia $\phi(X_1, X_2)$ ponderada por el signo de la asimetría dada en la definición 13.

Como parte de los elementos necesarios para establecer una regionalización en el aspecto heurístico se introducen los conceptos de rama y hoja de un MST.

Definición 14 *(Rama y hoja de un MST).*
Sea a_{ij} la arista determinada por los vértices v_i y v_j y $n(v)$ la paridad del vértice v. Si $n(v_i) > 1$ y $n(v_j) > 1$ se dice que la arista a_{ij} de un MST es una rama. Si alguno de los vértices tiene paridad 1 se denominará hoja.

Dependiendo del contexto, es pertinente establecer una ponderación (penalización) a la divergencia obtenida entre dos unidades de estudio. Por ejemplo, si se estudia la distribución del ingreso, aunque dos unidades tengan la misma distribución empírica, el ingreso per cápita puede ser más alto en una unidad que otra y eso incrementaría la divergencia. A esta discrepancia se le denominará *densidad*, en contraste con lo que se denominará *intensidad* (definición 15);

Definición 15 *(Intensidad y densidad).*

1. *La intensidad de un fenómeno expresado por la variable aleatoria X está definida por la estructura de su densidad de probabilidad F_X;*
2. *La densidad de la variable aleatoria X es una función de la medida relativa de la variable respecto a la dimensión, no necesariamente geográfica, del estrato de referencia.*

En la definición 15, el tamaño de un estrato puede tomar diversas formas; una de las más comunes sería ponderar por la población o bien el ingreso per cápita. Asimismo, el concepto de densidad se refiere en un sentido de concentración a diferencia del concepto homófono en el ámbito de probabilidad. En tal caso, donde haya ambigüedad sobre el uso del término se hará explícito a cual se refiere.

Para establecer una ponderación (penalización) se requiere definir una función $w(q_i, q_j)$

donde q_i y q_j son los tamaños de las unidades moleculares i y j respectivamente, de tal forma que tenga las siguientes propiedades:

1. $w(q_i, q_j) = 1$ si $q_i = q_j$;
2. $w(q_i, q_j) > 1$ si $q_i \neq q_j$;
3. $w(q_i, q_j)$ es creciente en $|q_i - q_j|$.

Para tal efecto, se establece la función $w(q_i, q_j)$ como sigue:

Definición 16 *(Función de peso).*
Si q_i y q_j son las dimensiones de los conjuntos A_1 y A_2 respectivamente, la función de peso o penalización por diferencia de sus magnitudes está dada por la expresión:

$$w(q_i, q_j) = \begin{cases} g\left(\frac{q_j}{q_i}\right), & q_i < q_j \\ g\left(\frac{q_i}{q_j}\right), & q_i \geq q_j \end{cases} \tag{5.21}$$
$$= g\left(\frac{q_j}{q_i}\right) I_{\{q_i < q_j\}}(q_i, q_j) + g\left(\frac{q_i}{q_j}\right) I_{\{q_i \geq q_j\}}(q_i, q_j),$$

donde $g(\cdot)$ es una función creciente.

5.7.1. Regionalización con criterio híbrido del caso unidimensional

La remoción de aristas en un *MST* es un problema de optimización en si mismo. Dado que la construcción del *MST* se basa en seleccionar, entre los nodos elegibles, aquel de mínima disimilitud. Es de esperarse que aquellas aristas de máxima disimilitud estén asociadas a fronteras entre regiones, particularmente si se considera la remoción de aquellas aristas que más reduzcan la suma de disimilitudes de todas las aristas. Tomando este criterio heurístico, se genera un *MST* mediante simulación y se remueven

las tres aristas de máxima disimilitud para establecer cuatro *sub-MST* que definen la regionalización correspondiente con cuatro componentes.

El proceso construcción del *MST* y remoción de aristas para generar una regionalización se muestra en la figura 5.7. En el diagrama G es un grafo, R es una regionalización, A es el conjunto de las aristas del *MST* y $\tilde{A}$ es el conjunto de aristas que se conectan con al menos dos aristas más, esto es, que no forman parte de la orilla del grafo.

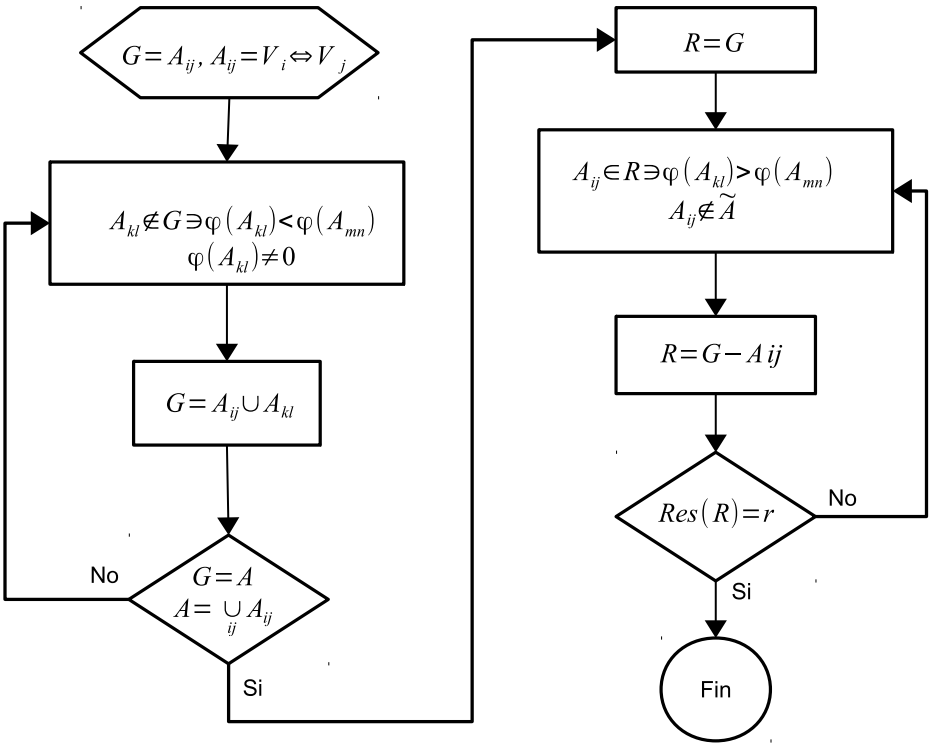
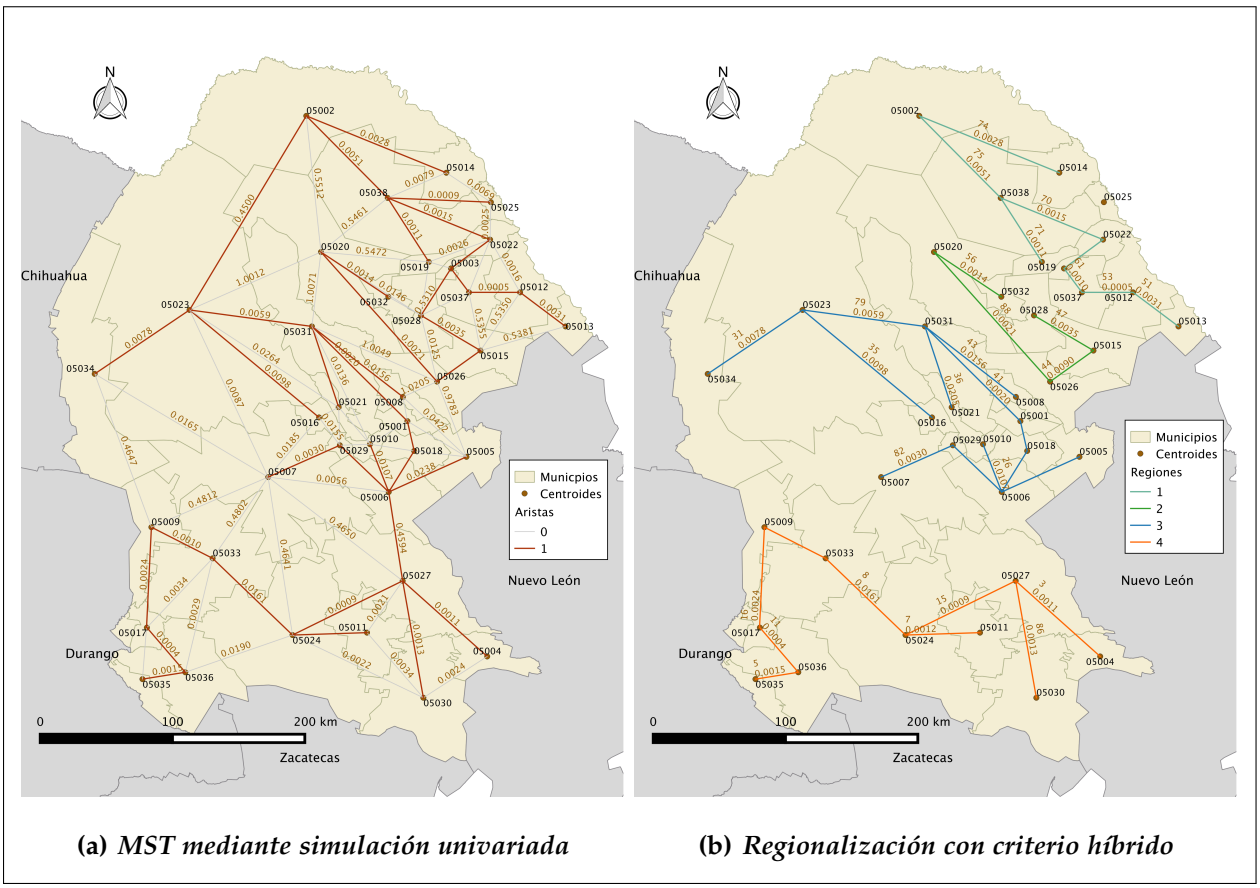


Figura 5.7: *Regionalización de un MST*

En la figura 5.8(a) se observa el *MST* que se obtiene mediante simulación, donde los valores en las aristas corresponden a la distancia geodésica $\phi(X_1, X_2)$.

Para establecer las regiones se eliminan las aristas de mayor disimilitud, las cuales representadas como un pares ordenados (*arista*, $\phi(X_1, X_2)$) son, en orden descendente, (0500305028, 0.5310), (0500605027, 0.4594) y (0500205023, 0.4500). Una vez retiradas estas aristas la regionalización obtenida se observa en la figura 5.8(b).

En la figura 5.9(a) se observa la regionalización obtenida con la remoción de las tres aristas de máxima divergencia. Esta regionalización se contrasta con la regiónalización

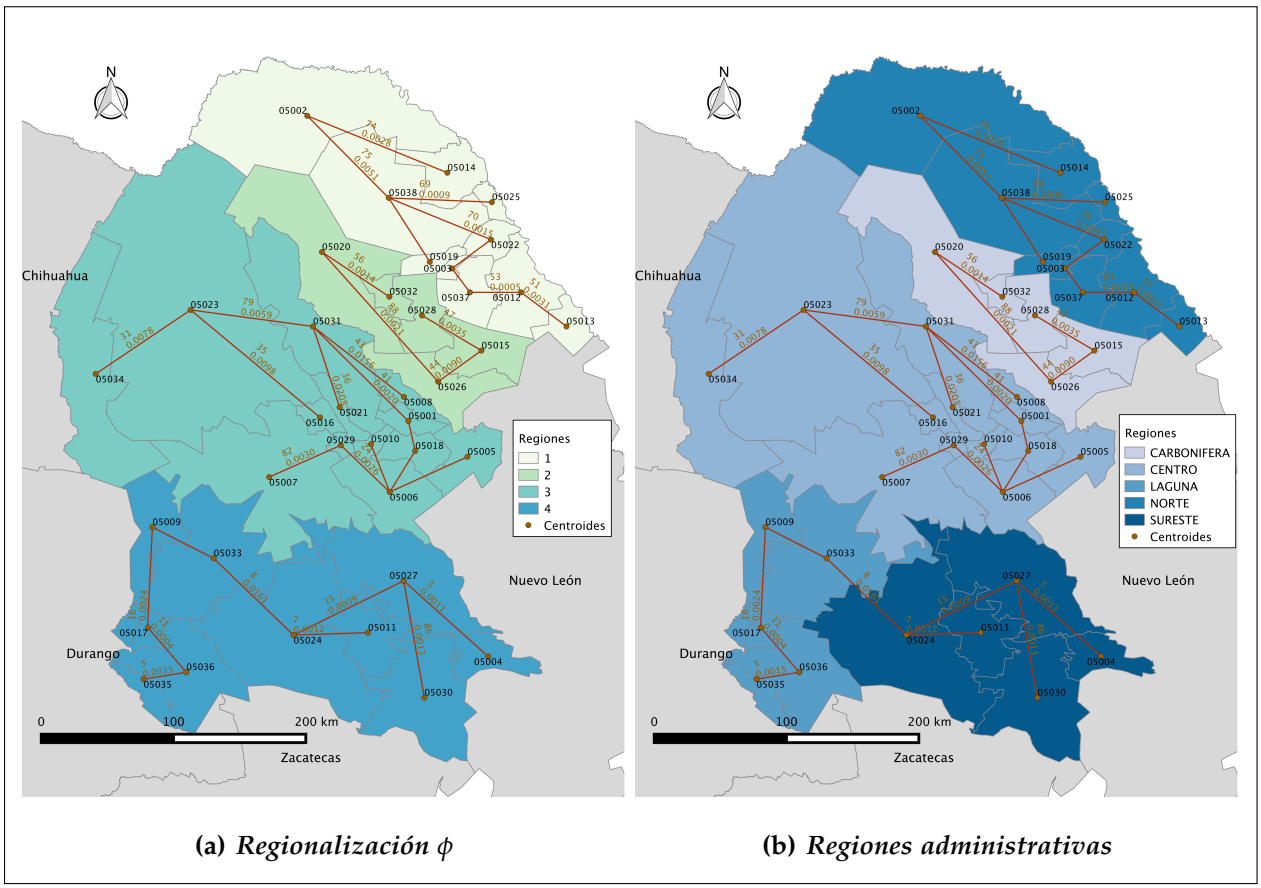


(a) ***MST mediante simulación univariada*** **(b)** ***Regionalización con criterio híbrido***

Figura 5.8: *MST y remoción heurística*

administrativa que se tomó como referencia para simular la distribución de probabilidad en los municipios correspondientes. Si la siguiente arista a remover es la de máxima divergencia que no corresponda a una arista en los extremos, esta es la que une a los municipios 05024 y 05033 cuyo valor es 0.0161 (figura 5.9(b)), entonces coincide con la partición inducida en la simulación.

En esta simulación, el método fue lo suficientemente sensible para detectar las regiones inducidas. En repeticiones de este proceso, los resultados son consistentes detectando en la mayoría de las veces las mismas regiones; esto es consistente ya que se trata de simulaciones con variaciones aleatorias donde ciertos municipios pueden, en distintos momentos, pertenecer a cualquiera de dos regiones contigüas.



(a) ***Regionalización*** ϕ **(b)** ***Regiones administrativas***

Figura 5.9: *Regionalización*

5.8. Alternativas de construcción de *MST* a mayor dimensión

El alcance planteado en este trabajo se centra en el caso unidimensional, sin embrago, se puede extender al caso multidimensional preservando la métrica y el método base.

Dos posibles variantes serían extender exactamente el mismo criterio aplicado en el caso unidimensional o bien dividir su aplicación en dos etapas: aplicar el criterio a cada uno de los componentes de los vectores aleatorios definidos por las variables de interés y después calcular alguna norma adecuada para estimar su divergencia.

En esta sección se realiza un ejercicio de lo que sería una primera fase de aplicación en el caso multivariado y se plantea una propuesta de reducción de dimensión mediante componentes principales.

5.8.1. Simulación de características en el caso multidimensional

Dado que la entropía desde la perspectiva de Shannon se aplica a una distribución de probabilidad, entonces esta es generalizable al caso multidimensional. Los ejemplos tratados han sido para el caso unidimensional y, dado que la estructura de un objeto espacial suele ser más compleja, entonces se generaliza al caso de múltiples atributos.

El problema de múltiples atributos puede abordarse de dos maneras:

1. Mediante la estimación de la distribución empírica de cada uno de los atributos del objeto espacial;

2. a través de la estimación de la distribución empírica multidimensional.

La primera tiene la ventaja de tratar con cada distribución por separado, formar un vector cuyas componentes sean la entropía de cada una de las distribuciones y obtener el vector de distancias *KL* de los atributos entre unidades espaciales mediante alguna norma establecida. Otra ventaja es que la cantidad de unidades observadas es menor que la requerida para el caso multivariado, reduciendo la cantidad de categorías en las que no se ubique ningún evento. La desventaja principal es que no toma en cuenta la correlación.

La segunda tiene como principal ventaja el considerar la correlación en forma implícita en la estimación de la distribución empírica. Además, se puede calcular la entropía de una una unidad espacial o bien la distancia *KL* multidimensional sin necesidad de incluir otra norma de distinta naturaleza a la de una medida de información.

Del caso multidimensional se observan dos desventajas principales: la primera es que se requieren más unidades de estudio observadas para tener una mejor estimación de la distribución y de las distancias entre unidades; la segunda es que las unidades observadas no siempre coinciden para dos atributos que se presentan en forma simultánea.

En la figura 5.10 se generan dos muestras aleatorias poblaciones normales bivariadas (X_1 y X_2) con los siguientes parámetros respectivamente:

$$n_1 = 1000, \quad \mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$n_2 = 1000, \quad \mu_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

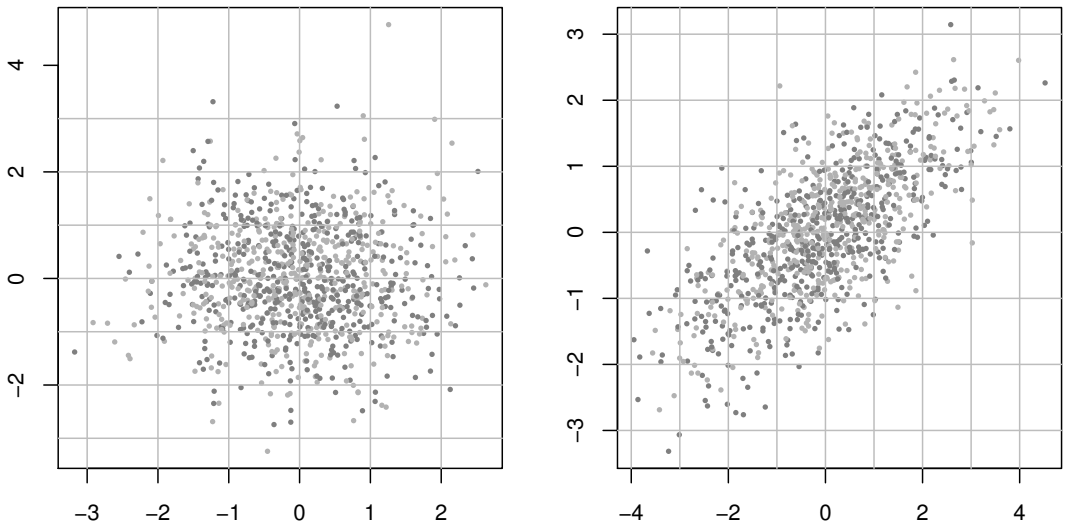


Figura 5.10: *Simulación de dos atributos (Normal)*

La primera de las distribuciones es con variables no correlacionadas con igual media y varianza; la segunda es a partir de variables correlacionadas y con medias iguales.

Si bien ambas distribuciones difieren significativamente en los parámetros, una distancia *KL* entre ambas no detectaría esto dado que al discretizar habría al menos una categoría sin valores, lo que llevaría a una disimilitud infinita. Sin embargo, la diferencia se puede estimar calculando la distancia de cada una de estas respecto a la distribución bivariada uniforme. Esta alternativa será útil en distribuciones multivariadas para resolver el problema de similitud al regionalizar.

Para este caso $KL(X_1, U) = 1.221$ y $KL(X_2, U) = 1.347$. Si bien no tenemos una referencia de la magnitud de la diferencia entre estas, estimaciones bootstrap revelarían la

significancia, si embargo para el objetivo de esta simulación no se realiza la prueba, lo cual se puede hacer una vez que se tenga una clasificación de las unidades de estudio.

5.8.2. Reducción de dimensión mediante componentes principales

Dado que componentes principales asigna una ponderación de acuerdo a la aportación de la variabilidad en los factores presentes, este recurso es viable para reducir la dimensión de variables. Por tanto, la metodología que se utiliza para el caso multidimensional en una regionalización de máxima información es la siguiente:

1. Obtener los componentes principales que aporten un porcentaje α dado de variabilidad;
2. estimar la distribución empírica para cada uno de los componentes;
3. dados dos unidades espaciales a comparar obtener la distancia definida en (5.20) componente a componente, formando el vector β de distancias;
4. calcular alguna norma adecuada de β.

Una de las principales ventajas de proceder de esa manera en el caso multivariado es que reduce sustancialmente el nivel de desagregación requerido en comparación a la distribución multivariada.

Por el alcance planteado en este trabajo, el método se valida mediante simulaciones y se clasifican los municipios de Coahuila en términos de las variables utilizadas para el cálculo de índice de marginación como variable univariada, dado que se utiliza la metodología de componentes principales en su cálculo.

5.9. Anotaciones

El método propuesto para el cálculo de la divergencia ϕ es cuantitativo e intensivo en datos. El desarrollo de esta medida realizó en tres etapas principales:

1. obtención de distancias y clasificación KL de unidades de estudio (municipios) respecto al universo de referencia (entidad) y en relación a una distribución uniforme;

2. corrección por signo de asimetría en distancia a la distribución uniforme para clasificación de unidades;

3. cálculo de la divergencia entre unidades de estudio para creación de MST y remoción de aristas para regionalización.

El criterio aplicado en la remoción de aristas con divergencia ϕ es en realidad un método híbrido al establecerse condiciones iniciales, a saber, que cada región tenga como mínimo dos unidades contiguas y se establece a priori la resolución de la partición (regionalización). Sin embargo, el peso del proceso de partición del universo de estudio se concentra en la medida de divergencia.

Aunque se dasarrolla aquí a detalle la aplicación de la divergencia información para el caso unidimensional como criterio para delimitar regiones, el método es generalizable para tipificar regiones mediante vectores de variables de interés.

CAPÍTULO 6

REGIONALIZACIÓN PARA INGRESO Y MARGINACIÓN

6.1. Introducción

Las asimetrías observadas en la población de un país o región queden en buena medida establecidas por la distribución de los ingresos. En el entendido de que una distribución perfectamente uniforme no es factible por diversas razones, su proximidad a este comportamiento aporta información relevante acerca del modelo económico de referencia. En este sentido, se hace necesario contar con una métrica que permita establecer la similitud entre regiones en relación a la estructura de la distribución del ingreso, entendida esta estructura en términos de la distribución de la probabilidad empírica observada.

La inequidad, entendida como lejanía a una distribución uniforme en el ingreso puede ser medida de diversas maneras, una de las más utilizadas en diferentes contextos es el índice de Gini, el cual se basa en la distribución acumulada establecida para alguna división de la población en cuantiles (Milton y Arnold, 1990), por ejemplo en quintiles o deciles. Otra medida poco utilizada en este contexto es la Información de Kullback-Leibler (Konishi y Kitagawa, 2008, p 31), interpretada como una medida de proximidad de una distribución de probabilidad a otra de referencia. Esta medida, considerada una seudo-distancia y denotada por I_{KL}, será tomada como referencia aquí para establecer un

criterio de disimilitud entre distribuciones empíricas.

Si bien el comportamiento de ingreso se fundamenta en la estructura de su distribución (cuantiles) existen otras medidas resumen de carácter multidimensional, tal es el caso del Indice de Marginación calculado por CONAPO (2010a), el cual incorpora nueve variables para el caso municipal y ocho para el caso de localidades. Existen varias limitaciones que tiene esta medida resumen, resaltando dos principalmente: la primera es que se toma como un valor agregado de todas las variables sin determinar la aportación individual de grupos de variables una vez que se calcula, aunque si está presente en los ponderadores de los componentes principales; la segunda es que se hace referencia a este índice como un valor individual pero no se ha prestado atención a su estructura probabilística. Analizar su comportamiento estructural aporta elementos relevantes para establecer diferencias regionales.

6.2. Contigüidad y georeferenciación

En el proceso de regionalización se requiere inicialmente definir algunos elementos base, entre estos se tiene a la georeferenciación de las unidades espaciales y la contigüidad entre estas. El primer elemento se obtiene de la base de datos SCINCE de INEGI y corresponde a las capas del estado de Coahuila hasta nivel manzana. Para el segundo hay que realizar un trabajo previo para crear la capa correspondiente.

Dado que el INEGI no considera el concepto de región como unidad espacial dentro de las entidades, la primera tarea es la de añadir a la capa de municipios la variable *NREG* que establece a cual región pertenece cada uno de estos. Las regiones son *CARBONÍFERA*, *CENTRO-DESIERTO*, *LAGUNA*, *NORTE* y *SURESTE*. Considerar estas regiones administrativas es fundamental para comparar entre una regionalización óptima y una establecida apriori por criterios ajenos a las variables de estudio.

En relación a la contigüidad, en lugar de considerar una matriz binaria que, en el caso de Coahuila, sería de orden 38×38 y de condición rala, se construye una matriz $\mathbf{A}$ de orden $m \times 2$, donde m es la cantidad de colindancias municipales. Para tal efecto se

procede como sigue:

1. Se asigna un índice numérico correspondiente al número de municipio para establecer una relación uno a uno $M_i \leftrightarrow CM_i$ donde M_i es el índice numérico y CM_i la correspondiente clave municipal;
2. se forma la matriz $\mathbf{A} = (\mathbf{u}_1|\mathbf{u}_2,) \in \Re^{m\times 2}$, donde $\mathbf{u}_1, \mathbf{u}_2 \in \Re^m$;
3. los vectores $\mathbf{u}_1$ y $\mathbf{u}_2$ cumplen la condición $\mathbf{u}_{1_i} < \mathbf{u}_{2_i}$;

Si cada par ordenado $(\mathbf{u}_{1_i}, \mathbf{u}_{2_i})$ corresponde al centroide municipal, entonces define una arista que representa la colindancia entre los municipios correspondientes. Para registrar esto, se crea una lista de claves formada por la concatenación de cadenas de caracteres de las claves municipales a 5 dígitos. Por ejemplo a la arista de colindancia de los municipios de Ramos Arizpe y Saltillo con claves *05027* y *05030* respectivamente, se le asigna la clave *0502705030*. Esta tabla será utilizada para agregarle atributos (distancias) y aplicar el algoritmo de regionalización. La delimitación regional y las aristas de colindancia de Coahuila se muestran en a figura 6.1.

6.3. Distribución del ingreso

6.3.1. **Distancias en distribución del ingreso**

¿Cómo construir una distancia entre distribuciones de ingreso?

Antes de contestar esta pregunta es necesario señalar la diferencia entre la el ingreso promedio y la distribución del ingreso. El ingreso promedio se enfoca a establecer un valor único de ingreso en una región, de tal manera que asignando a cada unidad de estudio esa cantidad, el ingreso total agregado sería el que se tiene de referencia. A diferencia de este, la distribución del ingreso adquiere una connotación multidimensional ya que establece una segmentación de cada región y se obtiene el ingreso agregado en cada uno de dichos segmentos. Entonces, la diferencia sustancial radica en que la primera no

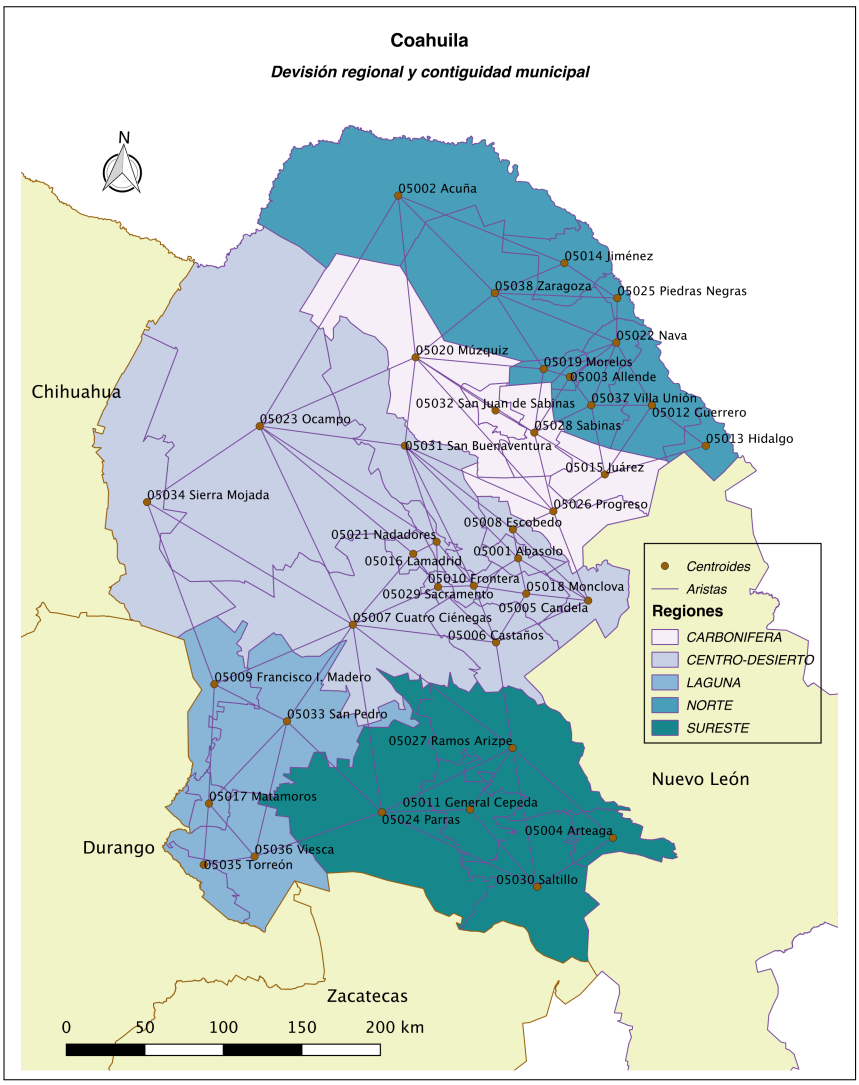


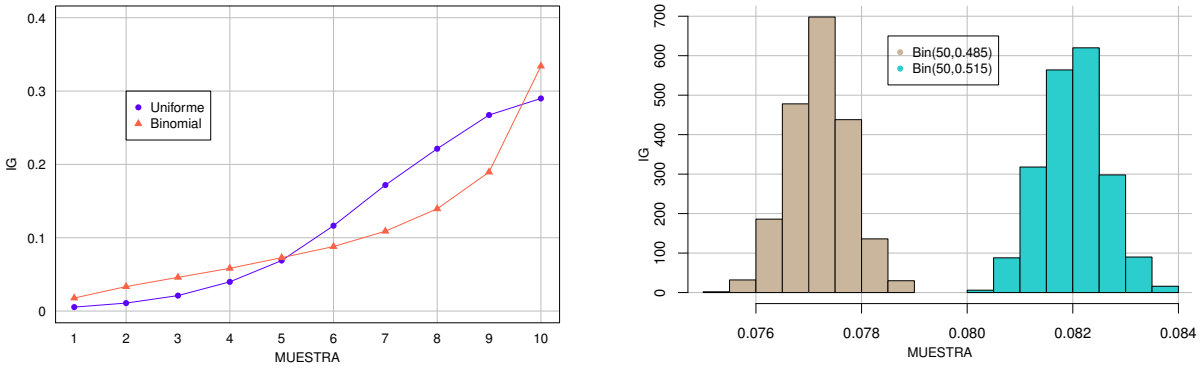
Figura 6.1: *Delimitación regional y contigüidad municipal en Coahuila*

considera la variabilidad mientras que la segunda si. Más aún, si se calculara una medida de variablilidad en el primer caso solo se estaría agregando cierta información adicional, más no al nivel de explicar su estructura probabiliítica en términos de su distribución.

Pretender medir la diferencia en distribución de ingreso en términos del ingreso promedio no responde cabalmente a la pregunta planteada. Sin embargo, considerar la segmentación de la población de referencia en términos del ingreso da la posibilidad de establecer su estructura y entonces trasladarlo a un contexto regional para realizar contrastes.

Una medida solo por el hecho de estar basada en la segmentación de una población no garantiza que mida efectivamente la estructura de la distribución de la variable de interés. Dada la relevancia y su uso extensivo se tomará como referencia el índice de Gini (IG) mediante el contraste de su comportamiento con las distribuciones uniforme y binomial. Para tal efecto, se harán dos comparaciones: entre distribución uniforme y binomial al cambiar sus parámetros y entre dos distribuciones binomiales con distinto parámetro.

La comparación se realiza mediante simulaciones para establecer el cambio del del IG al modificar parámetros de distribución uniforme y binomial; en cada caso se generan muestras aleatorias de tamaño 1000. Para la distribución uniforme se generan valores estableciendo el rango como $30, 30 + l$, donde $l = 2^k$ para $k = 0, 1, \ldots, 9$; esto significa que para $k = 0$ todos los datos tendrán el mismo valor y para $k = 9$ este será $(30, 542)$. Para la distribución binomial se genera con parámetros $(50, k)$, donde $k = 0.95, 0.85, \ldots, 0.05$. El resultado obtenido del IG para los 10 casos de cada distribución se muestra en la figura 6.2(a).



(a) ***Distribución Uniforme y Binomial***

(b) ***Distribución del IG en comportamiento Binomial***

Figura 6.2: *Comparación del Índice de Gini para distribución uniforme y binomial*

En la segunda comparación se obtiene la distribución empírica del IG utilizando la distribución binomial con valores del parámetro de probabilidad $p = 0.485$ y $p = 0.515$ simétricos respecto a 0.5; esto induce un sesgo hacia la izquierda y uno hacia la derecha, lo que a su vez genera un comportamiento de sesgo simétrico en la distribución empírica.

Hecho de esa manera se obtiene un cambio en el valor promedio, sin embargo se tiene la misma variabilidad. Se obtienen 1000 valores del *IG* para cada uno de los parámetros, dando como la distribución empírica de la figura 6.2(b).

Se puede observar en la figura 6.2(b) que aunque la distribución es similar, su valor medio es sensible a la dirección de la asimetría. Esto significa que aunque la *desigualdad* observada es similar el valor del *IG* es significativamente distinto. Los intervalos de confianza empíricos para el *IG* con $p = 0.485$ y $p = 0.515$ son $IC_{0.485} = (0.08086, 0.0832)$ y $IC_{0.515} = (0.07615, 0.0783)$ respectivamente. Dado que $IC_{0.485} \cap IC_{0.515} = \emptyset$ se rechaza la hipótesis de que tengan la misma media.

Para solventar el efecto de la distribución sobre el *IG* se utilizará la *Divergencia de Kullback-Leibler*, denotada por I_{KL}, como medida de disimilitud entre distribuciones (ver sección 5.5.2). En la figura 6.3 se muestran la funciones de distribución acumuladas del I_{KL} para las dos distribuciones binomiales simuladas, donde se observa prácticamente una coincidencia entre estas.

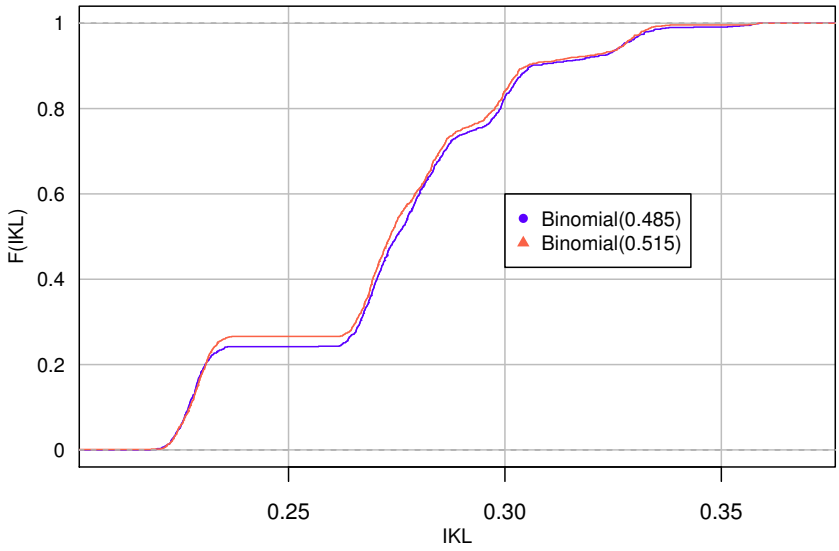


Figura 6.3: *Comparación I_{KL} para distribución binomial*

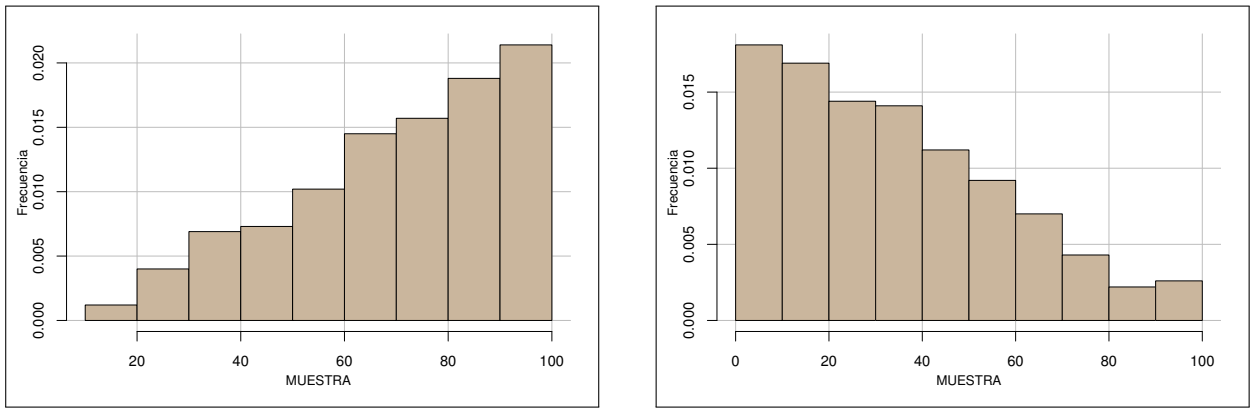
La diferencia observada en la distribución del *IG* en relación a la de *IKL* se enfatiza utilizando distribuciones donde se acentúa la asimetría hacia la derecha y hacia a la izquierda. Para tal efecto se simularon dos distribuciones triangulares mediante la expresiones dadas

en (6.1).

$$X_1 = 10 + \sqrt{8100U_q}, \quad X_2 = 100 - 10\sqrt{99U_2 - 1}, U_2 > \frac{1}{99} \tag{6.1}$$

donde U_1 y U_2 se distribuyen uniforme en $(0, 1)$.

Para cada una de las distribuciones se generaron 1000 datos. Los histogramas obtenidos para cada una de las distribuciones se muestran en las figuras 6.4(a) y 6.4(b) respectivamente.



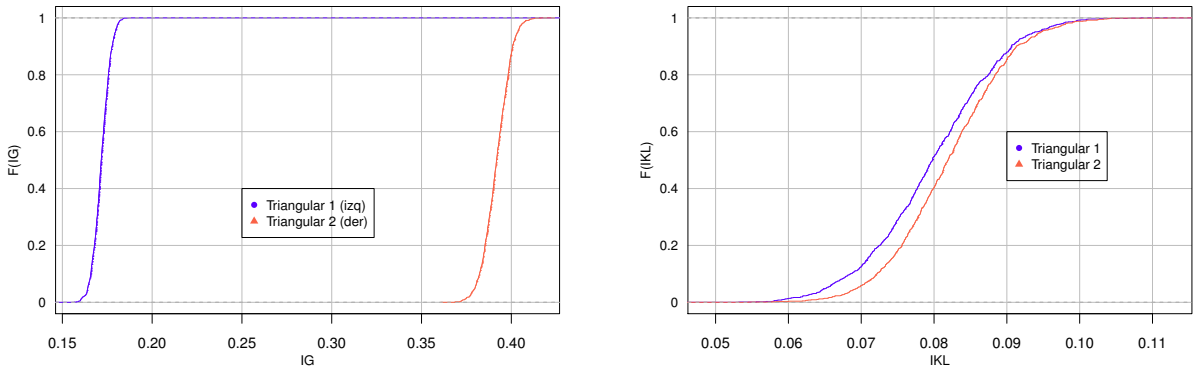
(a) ***Sesgo a la izquierda***

(b) ***Sesgo a la derecha***

Figura 6.4: *Distribución triangular*

Para muestras de 1000 datos simulados de X_1 y X_2 se generen a la vez 1000 repeticiones, calculando en cada una de estas el índice de *Gini* y la *Información de Kullback-Leibler*. Hecho esto, se obtiene la función de distribución empírica para cada una de las distribuciones. La figura 6.5(a) muestra el resultado para el *IG* y la figura 6.5(b) para la *IKL*.

Mientras que la distribución del *IG* se observa una diferencia significativa esto no ocurre en la *IKL*. Esto muestra que el primero es sensible a la asimetría, a diferencia del segundo que no lo es. Si consideramos que los datos corresponden a ingresos, nominalmente las dos distribuciones muestran la misma intensidad de desigualdad solo que con la asimetría en dirección contraria, sin embargo el *IG* las consideraría sustancialmente distintas, situación que no ocurre con la *IKL*. Esto establece una medida de intensidad absoluta, la cual será ponderada por la dirección de la asimetría para establecer distancias asimétricas.



(a) ***Índice de Gini*** **(b)** ***Información de Kullback-Leibler***

Figura 6.5: *Distribución acumulada del IG e IKL en distribución triangular*

6.3.2. Distribución de ingreso en Coahuila

Como referencia para el ingreso, se considerarán los ingresos por trabajo establecidos por *INEGI* como aquella *percepción monetaria que la población ocupada obtiene o recibe del (los) trabajo(s) que desempeñó en la semana de referencia. Se consideran los ingresos por concepto de ganancia, comisión, sueldo, salario, jornal, propina o cualquier otro devengado de su participación en alguna actividad económica. Los ingresos están calculados de forma mensual* (INEGI, 2011, p 41).

Los datos están dados a nivel de persona, a partir de los cuales se obtiene la distribución empírica para tipificar a los municipios y establecer las divergencias para su regionalización.

Se consideran tres alternativas para clasificar a los municipios en función de su distribución de ingreso: la primera es tomar como referencia la disimilitud con la distribución uniforme de ingresos; la segunda considerando a la distancia respecto a la distribución observada en todo el estado; la tercera es la aplicación de la divergencia ϕ para establecer distancias entre unidades de estudio..

A la distribución uniforme de ingresos se le denominará *ideal*, sin embargo no considerada como lo esperado en una distribución de ingreso óptima ya que esto, por diversos

factores, no aplica en el contexto de una realidad económica.

Para clasificar los municipios en términos del diferencial en distribución de ingreso se utilizarán microdatos del censo de población de 2010 del *INEGI* tomando como universo de referencia al estado de Coahuila, donde se establece su distancia a una distribución uniforme del ingreso mediante la Información de Kullback-Leibler. A partir de la medida I_{KL} se generan cinco categorías por rangos de distancia para clasificar a los 38 municipios del estado en términos de su proximidad (lejanía) a una distribución ideal.

A partir de los censos de población 2010 se recaba información relativa a los ingresos individuales mensuales por trabajo. La variable se denomina *INGTRMEN* y es codificada como *999998* para quienes tienen ingresos igual o mayores a $999,998.00 y como *999999* para aquellos valores no especificados.

Para tal efecto se tienen las siguientes consideraciones:

- Dado que el objetivo es establecer una medida de proximidad a una distribución uniforme de ingresos, se considerarán para el estudio aquellos valores que presenten ingresos menores a $999,998.00. Esto tiene sentido ya que no se trata de estimar el ingreso promedio, situación que si tendría un sesgo considerable si se eliminan aquellos individuos con ingresos mayores a la cota establecida.

- Como medida de aproximación a una distribución uniforme se utilizará la distancia o información de *Kullback-Leibler* (Konishi y Kitagawa, 2008, p. 28), donde se escribirá $I_{KL}(F, G)$ para denotar la medida de la distribución F a la distribución G. La distancia será tomada de manera unidireccional (ver apartado 5.5.2). Posteriormente se utilizará la medida de divergencia propuesta mediante la construcción de un *MST*.

- Para estimar la distribución de probabilidad se utilizan ocho categorías al dividir el rango de ingresos en ocho partes iguales. La decisión de tomar esta cantidad de categorías es con la finalidad de tener al *bit* como unidad de medida de información.

- La información está concentrada en formato nativo para el programa comercial

SPSS; esta es decodificada y procesada en el lenguaje *R* para obtener la medida que será considerada como distancia.

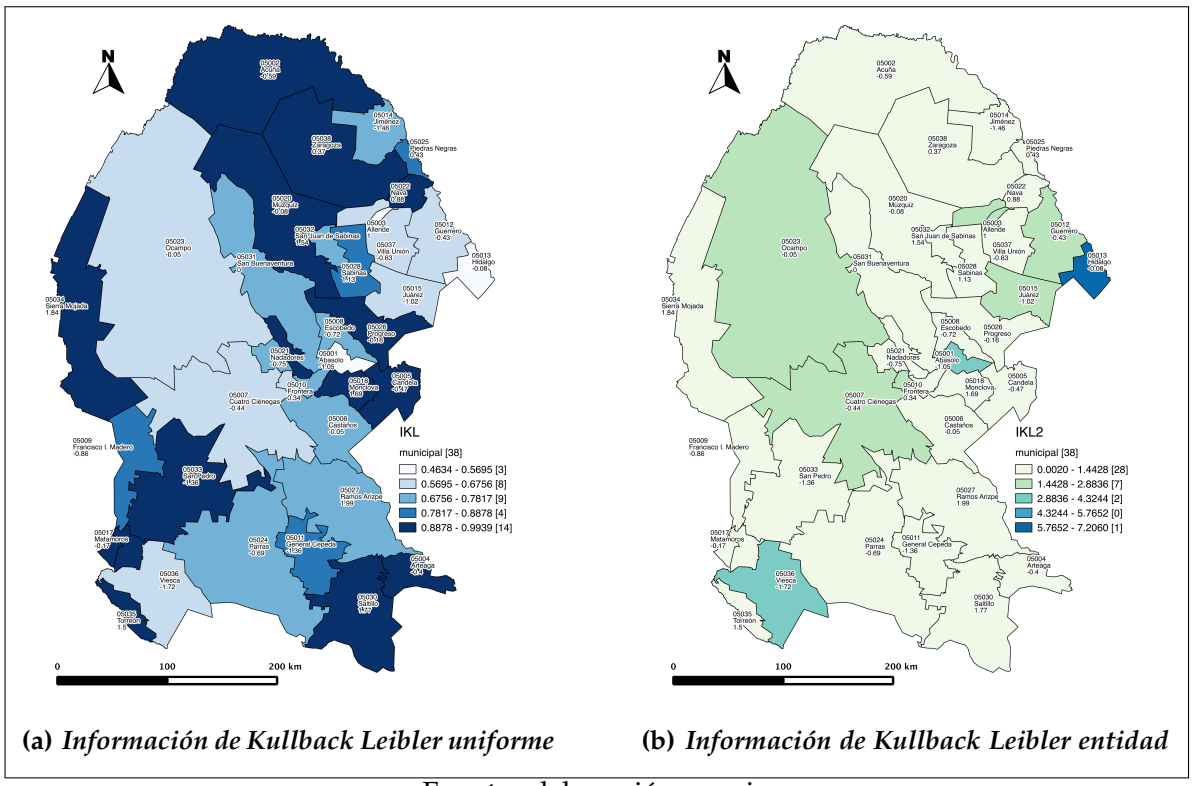
La distribución de frecuencias de ocho categorías en intervalos de igual extensión de población para Coahuila concentran en el cuadro C.2. Para fines de comparación se calculan cuatro medidas de distribución de ingreso: la información $I_{KL}(F, U)$, denotada como IKL, donde F es la distribución de ingreso en la muestra y U la distribución uniforme hipotética, la información $I_{KL}(F, E)$ denotada por $IKL2$, donde E es la distribución del ingreso en el estado, el índice de Gini G y el promedio de ingreso estandarizado denotado por M.

La cantidad de población en cada uno de los municipios juega un rol relevante en el cálculo de índices, esto debido a que la mayor cantidad se concentra en ciertas zonas. Los municipios de Saltillo y Torreón acumulan el 49.7% de la población total en el estado mientras que se tiene el 24.65% de la muestra.

La inclusión de cuatro medidas tiene como finalidad contrastar la distintas clasificaciones que se obtienen si dividimos en cinco categorías de igual extensión para cada una de estas.

En las figuras 6.6 (a) y (b) respectivamente, se observa la clasificación de municipios por medida de información *IKL* y la información *IKL2*. En las figuras 6.7 (a) y (b) respectivamente, se representan el *Índice de Gini G* y el ingreso promedio estandarizado *M*.

Si se jeraquiza de acuerdo a cada una de las medidas utilizadas aquí, es evidente que los municipios se ubican en distinta posición para cada una de las medidas utilizadas. Una ventaja de utilizar I_{KL} sobre *Gini* es que en la construcción de regiones podemos obtener las distancias estructurales entre municipios contigüos mientras que que con *Gini* o el valor promedio se tendrían medidas resumen cuya distancia sería la diferencia de valores individuales, dicho de otra manera, la primera se obtiene como una medida resumen de las diferencias, mientras que las otras dos corresponderían a la diferencia de dos medidas resumen.



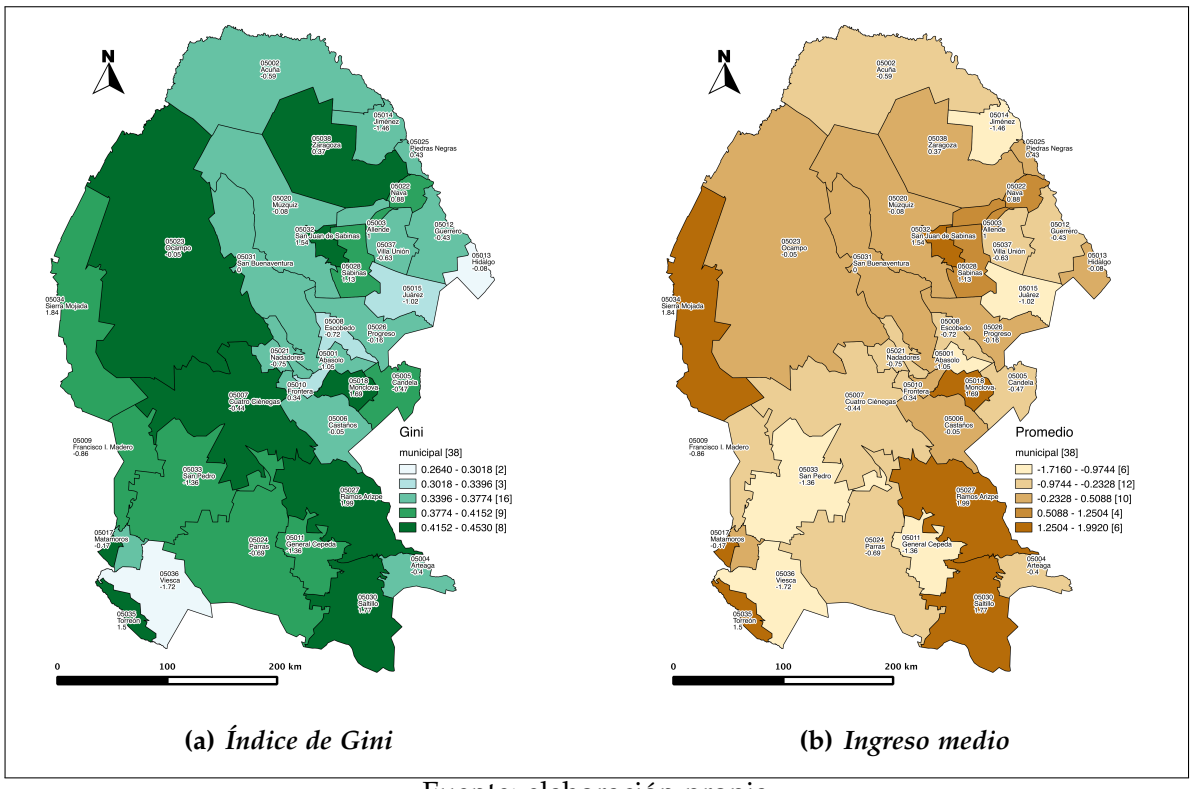
(a) ***Información de Kullback Leibler uniforme*** **(b)** ***Información de Kullback Leibler entidad***

Fuente: elaboración propia.

Figura 6.6: *Categorías por intensidad de ingresos en Coahuila mediante I_{KL} respecto a una distribución uniforme y a la entidad*

Más aún, aquí I_{KL} se calcula comparando la distribución de ingreso de cada municipio contra la distribución uniforme, metodología que servirá también para obtener dismilitudes entre grupos de municipios, esto es, entre regiones. En contraste con I_{KL} calculado como distancia a la distribución de ingreso en la entidad, se puede notar más uniformidad en la mayoría de los municipios; esto es razonable pues reflejan en cierta medida el comportamiento de la distribución general. En este sentido, comparar contra una distribución teórica de referencia (no necesariamente la uniforme) da una mejor perspectiva de las diferencias observadas entre municipios.

Se debe tener cuidado antes de emitir algún juicio asociado acerca de las diferencias observadas entre los municipios y entre medidas por dos razones principales: la primera es que los datos provienen de una muestra dentro de un marco cuasi-experimental, esto es, corresponden a aquellos individuos que contestaron al la pregunta; la segunda es



(a) ***Índice de Gini*** **(b)** ***Ingreso medio***

Fuente: elaboración propia.

Figura 6.7: *Categorías por intensidad de ingresos en Coahuila mediante I_{KL}, Índice de Gini e ingreso medio*

que los resultados están dados en términos de la distribución, desigualdad e ingreso percápita estimado, los cuales aportan información sustancialmente distinta. Dependiendo del contexto de interés tomará relevancia una u otra medida, particularmente aquí interesa comparar la estructura expresada en función de la distribución del probabilidad empírica del ingreso, lo que se hace a través de I_{KL}. Más aún, se debe considerar que hay un sesgo ya que los ingresos se truncan; una posible alternativa, que no se explorará en este estudio, es realizar estimación considerando que se cuenta con una muestra de datos censurados.

Una distribución del ingreso más cercana a la ideal teórica, esto es, tener una distribución uniforme, no necesariamente se traduce en una mejor condición. Esta es información relevante en lo que se refiere a la composición estructural al segmentar los ingresos en ocho categorías en la muestra de referencia. Por ejemplo, se observa en el cuadro C.2 que el valor mínimo de *IKL* corresponde al municipio de Hidalgo y el máximo a Matamoros, esto es, el municipio que más se acerca a una distribución uniforme y el que más se aleja

respectivamente. En el caso de *IKL2*, donde se toma como referencia a la distribución de ingreso en la entidad, la relación es inversa a la obtenida en *IKL*, esto es, aquí el municipio de Hidalgo tiene un valor mayor que el de Matamoros.

Se destacan algunas de las características relevantes que tiene el I_{KL} en relación a otras medidas como *Gini* o valor promedio:

- establece una medida de disimilitud basada en la distribución de frecuencias (ingreso) en lugar de una distancia basada en una medida resumen. Esto es, obtiene una medida resumen pero después de comparar una a una las categorías de dos distribuciones;
- a diferencia del Índice de Gini, la I_{KL} no es sensible a la dirección de la asimetría respecto a la distribución de referencia, en este caso la uniforme;
- tomada como referencia la distribución uniforme nos da una distancia comparativa de municipios inclusive a otros que no sean de la propia entidad;
- con una distribución base de comparación, se tendrá un mejor contraste de las medidas de distribución entre municipios, situación que no se ocurre si se toma como referencia a la entidad.

Finalmente, se considera establecer la distancia (disimilitud) entre unidades de estudio en términos de divergencia de información. Aunque utiliza como recurso las divergencias respecto a una distribución uniforme, calculada de esta manera amplía el espectro de aplicación aún contando con pocos datos de la variable de interés.

Regionalización del ingreso con divergencia

La divergencia $\phi(X, Y)$ entre los municipios X y Y aplicada en forma directa determina la diferencia de las estructuras de la distribución de probabilidad de estos. Esto no aporta información acerca de las diferencias en relación al ingreso medio, sino a como se distribuye el total de ingresos entre la población. Sin embargo, puesto en una escala

categórica común basada en el universo de referencia, las asimetrías de las distribuciones ponderan de manera implícita la divergencia obtenida.

Dado que la regionalización se obtiene a partir de la distribución del logaritmo del ingreso, para compensar esto en la función (5.21) se utilizará la función $g(x) = e^x$ para establecer la ponderación $w(q_i, q_j)$:

$$w(q_i, q_j) = e^{\frac{q_j}{q_i}-1} I_{\{q_i<q_j\}}(q_i, q_j) + e^{\frac{q_i}{q_j}-1} I_{\{q_i\geq q_j\}}(q_i, q_j), \tag{6.2}$$

Una vez aplicada la penalización dada en (6.2) se obtiene el *MST*. A partir de este, la eliminación de aristas para establecer las regiones correspondientes se realiza considerando conjuntos de cuando menos dos municipios (no aislar hojas del *MST*) para una resolución de 6 regiones en este caso (figura 6.8).

En la parte superior izquierda de la figura 6.8 se ubica el *espectro* de cada una de las regiones. Este espectro es el kernel de la distribución empírica del logaritmo del ingreso en cada una de las regiones generadas. Si bien gráficamente no es visible una diferencia sustancial, la partición realizada es óptima en el sentido de mínima divergencia intraregional y máxima interregional.

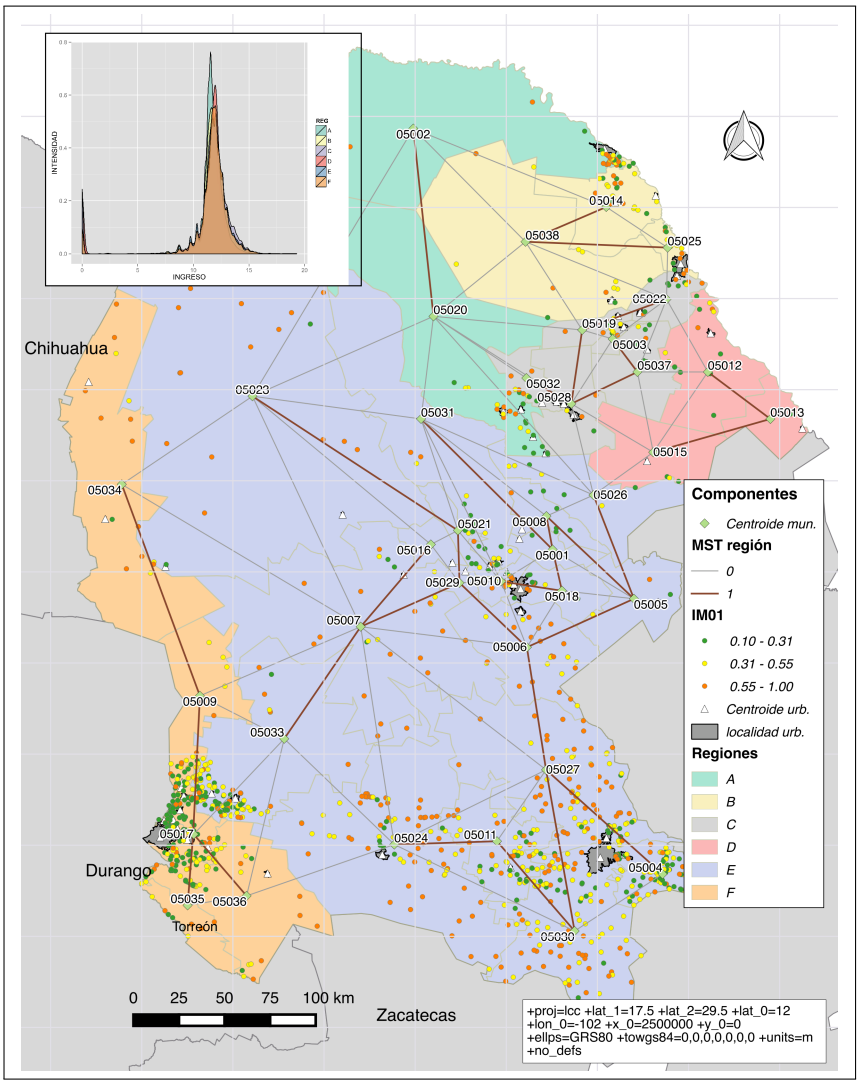


Figura 6.8: *Regionalización por intensidad de ingreso*

6.4. Marginación

El índice de marginación, al que se denotará como *IM*, es un indicador que pretende medir el grado de carencias que padece la población. A nivel estatal y municipal se agrupan en cuatro dimensiones: educación, vivienda, distribución de la población e ingresos por trabajo. Estas dimensiones se dividen a su vez en nueve formas de exclusión (CONAPO, 2010b, p 11-14):

- Educación
  - analfabetismo;
  - población sin primaria completa.
- Vivienda
  - sin drenaje ni excusado;
  - sin energía eléctrica;
  - sin agua entubada;
  - con algún nivel de hacinamiento;
  - con piso de tierra.
- Distribución de la población
  - localidades con menos 5000 habitantes.
- Ingresos monetarios
  - población ocupada que percibe menos de dos salarios mínimos.

En el cálculo de este índice se aplica la metodología de componentes principales (Jobson, 1991, p 345), de tal forma que a partir de los dos primeros que acumulan más variabilidad se proyecta cada unidad de estudio sobre el primero de ellos (CONAPO, 2010b, p. 322-324). Este proceso implica que los datos, medidos como rezagos en porcentajes, son sometidos

a un proceso de estandarización ubicando a la mayoría de los índices de marginación en el rango de -3 a 3, aunque nominalmente no es acotado el rango de valores posibles. Para el caso del cálculo a nivel localidad la variable de localidades de menos de 5000 habitantes no aplica.

6.4.1. Características del Índice de Marginación

Se destacan tres características principales del índice de marginación: es una medida resumen, no tiene una escala acotada universal y no es comparable en el tiempo (CONAPO, 2013, p 12). Esta última al menos no bajo la misma metodología de componentes principales.

Como medida resumen el *IM* condensa la información de las variables involucradas en un solo valor, lo que representa una desventaja si se quisiera determinar cuánto aporta cada uno de los factores involucrados en su cálculo. Estos factores se pueden agrupar en bloques a los que llamaremos ejes; en este caso son *educación*, *vivienda*, *distribución de la población* e *ingreso.*

En la metodología de cálculo del *IM* este pasa por un proceso de estandarización, tanto en la construcción de la matriz de correlación como en la proyección final sobre el el componente de mayor variabilidad. Esto tiene como efecto el que no esté acotado aunque la mayoría de los valores se distribuyan en el intervalo $(-3, 3)$. La distribución de probabilidad aproximada a la normal estándar del *IM* se aprovechará para acotarlo.

6.4.2. Escala acotada del *IM*

La escala natural que deriva de la metodología en el cálculo del índice de marginación dificulta su lectura, particularmente cuando se trata de comparar distintas unidades geográficas. Este efecto ocurre principalmente porque no es acotada ni lineal en la escala por el proceso de estandarización al que se ven sometidos los datos. Es precisamente este proceso de estandarización el que facilita un cambio de escala de tal manera que se pueda acotar.

El IM sigue aproximadamente una distribución normal estándar. La figura 6.9 muestra el índice de marginación en Coahuila para el 2000 ($IM00$), para 2010 ($IM10$), para 2000 y 2010 en forma conjunta ($IM0010$) y la función de densidad acumulada de la distribución normal estándar ($FDAZ$).

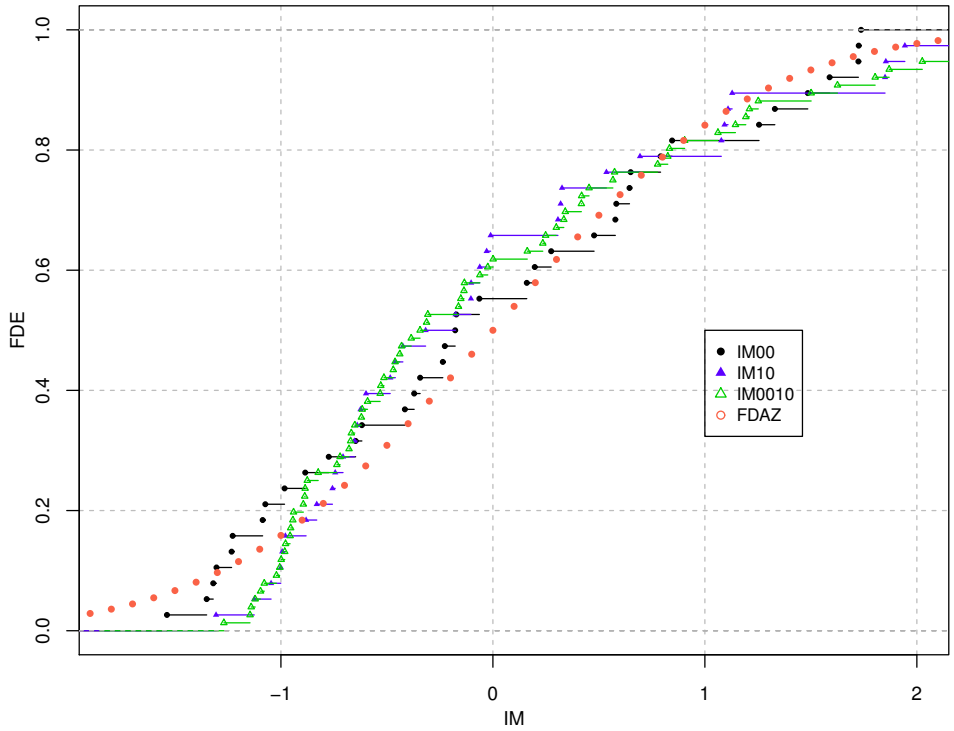


Figura 6.9: *Función empírica del IM y función de distribución acumulada normal*

Considerando el comportamiento del IM es factible hacer una transformación que permita preservar el orden, lo acote y que además que sea asintótico en en los valores mínimo y máximo de rezago. Para tal efecto se propone la transformación monótona del IM dada por la expresión (6.3).

$$IM01 = T(IM) = \int_{-\infty}^{IM} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du \tag{6.3}$$

Se observa en (6.3) que, dado el comportamiento aproximadamente normal, el $IM01$ corresponde a la probabilidad acumulada si se considera al IM como un cuantil de la distribución normal estándar. Una característica a destacar es la tendencia asintótica, ya

que en la realidad una unidad de estudio difícilmente se ubicaría en condiciones límite de rezago, esto es, de tener cero rezago o rezago total en todas las variables utilizadas para el cálculo del IM.

Por la naturaleza de la transformación la pendiente del $IM01$ en valores cercanos a 0.5 es más pronunciada y menos conforme el valor se acerca a 0 o 1; esto refleja un hecho que se observa en la realidad, a saber, que unidades geográficas con marginación cercana a 0 requieren una reducción significativa en los rezagos para observar un cambio significativo en el valor del índice o bien, en contraste, las unidades con marginación cercana a 1 se mostrarán una mayor reducción del valor índice. Esto se puede comprobar obteniendo la derivada del $IM01$ respecto a IM en la expresión (6.3) mediante la aplicación de la regla de *Leibnitz* (teorema 4, apéndice A) de derivación bajo el operador integral con el resultado siguiente:

$$\frac{dIM01}{dIM} = \frac{dIM01}{dIM} \int_{-\infty}^{IM} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}IM^2} \tag{6.4}$$

La figura 6.10 muestra gráficamente el efecto de reducción del IM en dos extremos de la escala. Si dos unidades económicas presentan valores de IM de -2 y 2 respectivamente, la reducción en una unidad en cada una de ellas es representada por la longitud del intervalo I_0 y I_1 respectivamente. Se puede demostrar que $|I_1| = 0.1359$ y $|I_0| = 0.0214$. Es claro también que la máxima sensibilidad del $IM01$ al cambio en el IM se tiene en una vecindad de $IM = 0$ donde se presenta la mayor pendiente.

6.5. Regionalización con IM rural

El índice de marginación, como un índice agregado, aporta información parcial de la magnitud absoluta y dificulta establecer diferencias entre municipios, por tanto no es una medida lo suficientemente fina para aglutinar grupos y formar regiones. Para hacer esto se aplica la divergencia ϕ mediante la construcción de un *MST* bajo condición de contigüidad.

Para la estimación de la distribución empírica que servirá de elemento de comparación

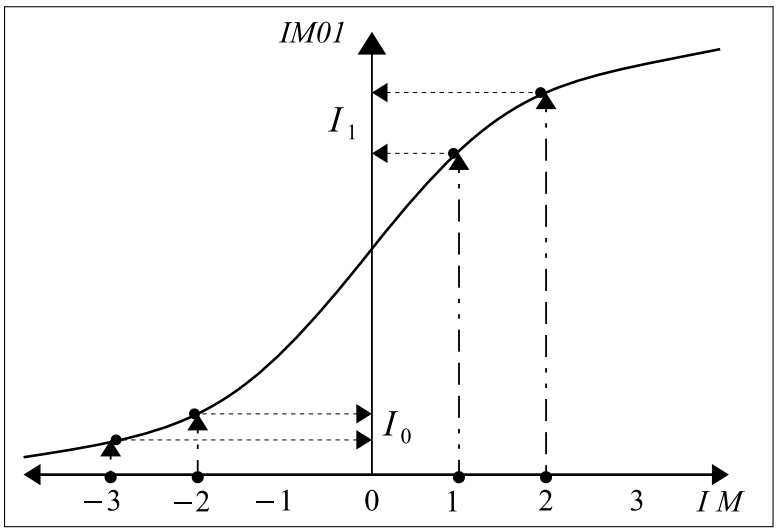


Figura 6.10: *Reducción del IM01 en función del IM*

entre municipios, se utilizará el conjunto de índices de marginación de las localidades rurales[1]. Se utilizan estas localidades para tener una cobertura geográfica del municipio y no solo concentraciones como el caso de las zonas urbanas, el cual es un caso de estudio por si solo. Esto permite tener un panorama más amplio de las comunidades y zonas donde se agudizan más las carencias consideradas en el índice.

Se reproduce el cálculo del índice de marginación calculado por *CONAPO*, tomando directamente las variables del censo 2010 de *INEGI*. Dado que el objetivo es contrastar las distintas localidades en Coahuila, el cálculo se realiza en forma relativa a esta entidad. Este mecanismo tiene la ventaja de que se puede extender a zonas preestablecidas, inclusive aquellas que combinen municipios contiguos de distintas entidades y, por tanto, aplicable al país completo.

Una vez calculados los índices de marginación *IM01* se estima la distribución empírica con 8 categorías, se obtiene la divergencia ϕ y a partir de esto se genera el *MST*. Una vez aplicado el criterio propuesto para la eliminación de aristas de máxima divergencia con una resolución de 5 regiones, se obtiene la segmentación representada en la figura 6.11. Al igual que en la distribución del ingreso, en la parte superior izquierda se muestra

[1]A diferencia del *IM* municipal, aquí no se considera la variable de poblaciones de menos de 5000 habitantes ni la estimación del ingreso basado en si se tiene o no refrigerador.

el espectro (kernel) de las distribuciones de probabilidad empíricas de cada una de las regiones.

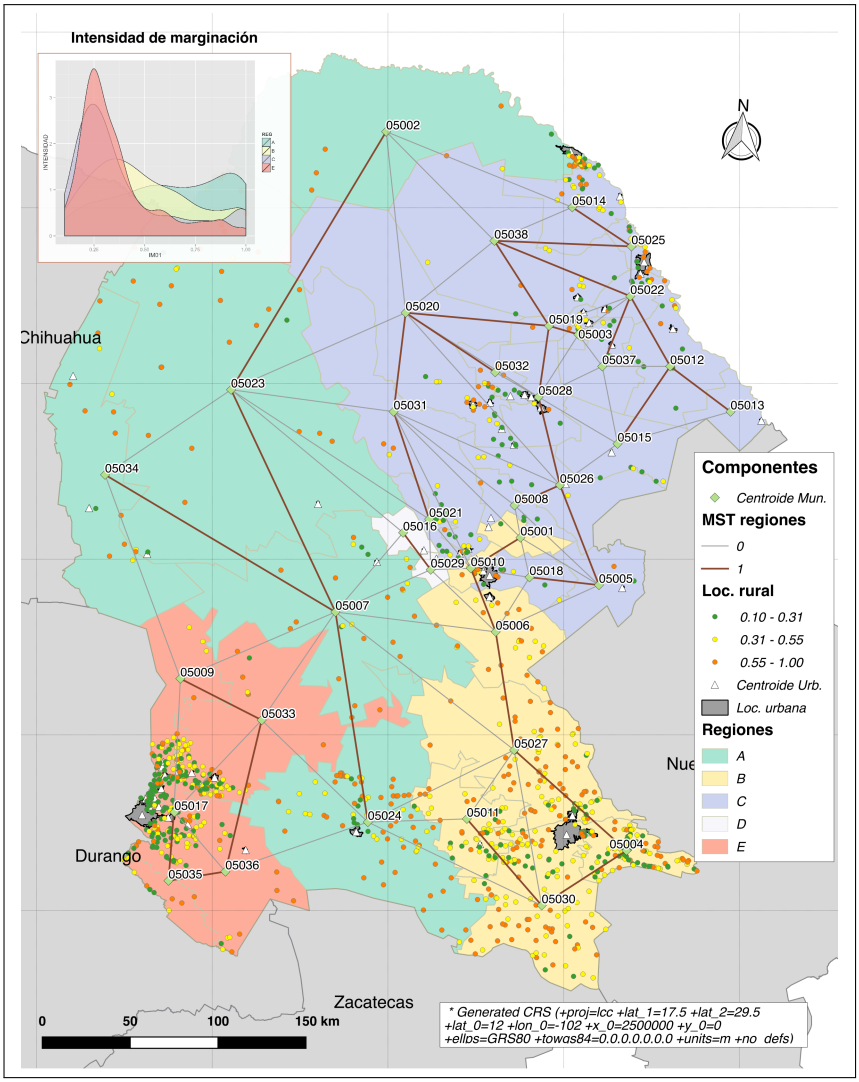


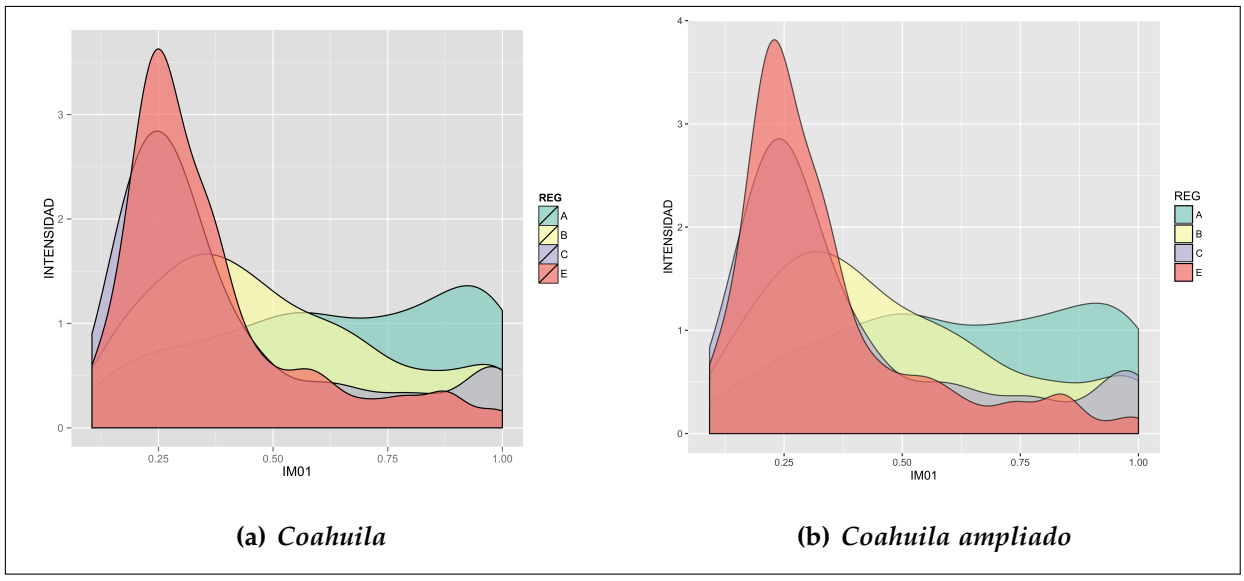
Figura 6.11: *Regionalización por intensidad de Marginación rural*

A diferencia del espectro para el ingreso, en el caso del índice de marginación, las diferencias observadas son claras, mostrando inclusive cambio de la dirección de la asimetría entre regiones. Si se contrasta esta partición con las regiones administrativas, se puede notar que dos de estas se aproximan a la región laguna y sureste.

6.5.1. Efecto de municipios contiguos de entidades vecinas

Dado que Coahuila no es un espacio aislado de lo que pasa con las entidades vecinas, es pertinente integrar a municipios colindantes de estas entidades (Chihuahua, Durango, Zacatecas y Nuevo León) y calcular el índice de marginación integrando a todas las localidades rurales correspondientes.

Dado que ya se tiene la regionalización de mínima divergencia de información tomando a Coahuila como universo, basta comparar el conjunto de kernels obtenidos de esta forma contra el obtenido en Coahuila pero calculado con todas las localidades rurales de los municipios colindantes de las entidades vecinas. A este conjunto se le denominará Coahuila ampliado. Las figuras 6.12(a) y 6.12(b) muestran los respectivos kernels.



(a) ***Coahuila*** **(b)** ***Coahuila ampliado***

Figura 6.12: *Kernel de marginación rural para regiones de Coahuila*

El conjunto de kernels obtenido en espacio de Coahuila ampliado, se puede considerar como condicional ante la presencia de localidades rurales externas. Esto da la posibilidad de establecer en que medida influye el integrarlas. El resultado obtenido es que los kernel permanecen sin cambio significativo (esto se puede validar calculando la entropía de cada uno).

La estructura de rezagos en los municipios de Coahuila no cambia al integrar los muni-

cipios externos. Por tanto, el orden del índice de marginación obtenido entre municipios no se ve alterado, de ahí que el kernel refleja la misma estructura de la distribución de probabilidad empírica. Dicho de otra manera, los cuantiles de los índices de marginación de las localidades no cambian su orden relativo.

6.6. Anotaciones

Por la naturaleza de la variable de estudio, los espectros (kernel) de las regiones obtenidas por ingresos son menos contrastantes entre si que los obtenidos a partir del índice marginación.

A diferencia del *MST* para ingreso, en el caso de marginación los municipios de Torreón, Matamoros y Francisco I. Madero no se conectan con el municipio de Sierra Mojada.

Dado que la divergencia es una medida definida en los reales, difícilmente se encontrarán dos aristas que presenten el mismo valor, hecho que posibilita escoger una resolución arbitraria, limitada solamente por la cantidad de unidades de estudio y las condiciones establecidas de acuerdo al contexto de estudio.

La inclusión de localidades rurales de municipios colindantes de las entidades vecinas no altera el conjunto de kernels obtenido sin esta condición. Esto refleja, en primera instancia, que los municipios externos colindantes tienen un comportamiento similar con los de Coahuila.

Una extensión al caso multivariado de la regionalización por ingreso y marginación en forma individual, es obtener las regiones considerando la distribución empírica del vector formado por ambas. Más aún, se puede considerar probar la hipótesis de propensión relativa a formar regiones mediante la función de riesgo dada en la expresión (5.3).

CAPÍTULO 7

ANÁLISIS DE RESULTADOS

El capítulo se divide en cinco apartados: el primero se centra en los resultados obtenidos mediante simulaciones y la sensibilidad de la medida de divergencia propuesta; en el segundo se establece una medida de la eficiencia de una regionalización basada en entropía, la cual es fundamental para verificar la calidad de las regionalizaciones; el tercero se enfoca en los resultados obtenidos de la regionalización por ingreso; en el cuarto se argumentan los resultados derivados de la regionalización tomando como referencia el índice de marginación; el quinto se centra en la comparación de las regionalizaciones de ingreso y marginación.

7.1. Resultados en simulaciones

La regionalización obtenida en la simulación realizada en el apartado 5.6.1 mostró la sensibilidad de la medida divergencia ϕ para detectar las regiones inducidas, las cuales se hicieron coincidir con las correspondientes administrativas establecidas en el estado de Coahuila.

La sensibilidad del método para identificar regiones es independiente de la elección de municipios que las conforman. Esto se atribuye al hecho de que en esencia lo que se com-

para son distribuciones de probabilidad de la variable de estudio y no una identificación de índole geográfica. Más aún, por construcción, la partición obtenida deberá tener una divergencia acumulada mínima en relación a todas las particiones posibles de la misma resolución bajo la condición de contigüidad.

Si bien se aplica el criterio de no eliminación de aristas que sean hoja en el *MST*, este mecanismo híbrido no necesariamente incrementa en forma significativa la divergencia interregional, su aplicación estricta dependería del contexto del espacio de estudio, particularmente si se identifican unidades de gran dimensión en relación a las unidades de menor dimensión que la conforman y el espacio geográfico que ocupan.

7.2. Eficiencia de la regionalización

La eficiencia de una regionalización se plantea en términos de un criterio de comparación adecuado. El criterio aplicado puede cambiar en términos de la función objetivo establecida. En el caso de la regionalización basada en mínima divergencia una medida natural es la entropía media (definición 17).

Definición 17 *Sea $\Omega = \{x_1, x_2, \ldots, x_n\}$ donde las x_i son variables aleatorias. Si $G = \{G_1, G_2, \ldots, G_k\}$ es una partición dada de Ω, se define la* entropía media *de la partición G como una función del vector $\mathbf{x}' = (x_1, x_2, \ldots, x_n)$:*

$$E_m(\mathbf{x}) = \frac{E_\Omega(\mathbf{x}) - \frac{1}{k}\sum_{i=1}^{k} E_{G_i}(\mathbf{x})}{E_\Omega(\mathbf{x})} \tag{7.1}$$

donde $E(\cdot)$ es la entropía de Shannon.

La cantidad dada en la expresión 7.1 determina la proporción de entropía en que difiere un conjunto de referencia en relación a la entropía promedio de la partición de referencia. Si la partición se asemeja a una selección aleatoria entonces la cantidad $E_m(\mathbf{x})$ tenderá a cero dado que cada subconjunto se parecerá en distribución al de referencia Ω. Por otro lado, si los grupos de la partición difieren en la distribución de probabilidad empírica

de las x_i entonces la cantidad será significativamente mayor que cero; esto significa que existen evidencia para rechazar la hipótesis $\mathcal{H}_0 : AEC$.

7.3. La hipótesis AEC

Si en un conjunto de referencia existen grupos (partición) que difieren entre si en su distribución empírica, entonces se presenta heterogeneidad y, por tanto, una segmentación del conjunto aporta más información. En el contexto de una aplicación este resultado equivale a decir que se identifican regiones, es decir, una regionalización es plausible, lo que significa que se rechazaría la hipótesis de homogeneidad en la distribución de probabilidad de las variables de estudio.

Es fundamental establecer que la presencia de AEC no se asocia estrictamente a una distribución de probabilidad uniforme, sino a la diferencia entre las distribuciones de probabilidad de las regiones identificadas y la observada en el universo de referencia.

Los experimentos in silico centrados en simulaciones facilitan la construcción de regiones ad-hoc. Esto significa que, en términos de probabilidad, se identifican grupos afines en la variable de estudio. En la inducción hecha en este sentido, las regiones se identificaron con quintiles de la distribución normal estándar. A través de la medida de divergencia ϕ y mediante la construcción del MST asociado con una resolución de cuatro regiones, el método detectó las regiones inducidas que previamente se sabe no cumplen con la hipótesis $\mathcal{H}_0 : AEC$. Mediante un criterio híbrido, la remoción de una arista adicional nos lleva a la identificación de las cinco regiones administrativas consideradas en el estudio.

La función de eficiencia en regionalización es la entropía media dada en la definición 17. Dado que el contraste de la hipótesis es establecido mediante la información que aportan los datos, la distribución empírica de E_m corresponde a la hipótesis nula o contrafactual de la existencia de regiones. Esta construcción se realiza mediante permutaciones de las mediciones entre unidades de estudio.

Para aplicar el criterio de permutaciones se toman todas las mediciones y se reordenan

para asignarlas de nuevo a las unidades de estudio manteniendo la dimensión asociada a cada una de estas. Este proceso se repite y en cada caso se calcula la entropía media. A partir del vector resultante de este proceso se obtiene la distribución empírica de referencia para ubicar el cuantil de la E_m de mínima divergencia.

Teorema 3 *(Convergencia de la distribución empírica regional)*
La distribución empírica en las regiones obtenida mediante permutaciones converge a la distribución del universo de referencia.
Demostración
La demostración se sigue directamente ya que cada permutación equivale a una muestra aleatoria del universo de referencia.

Del teorema 3 se desprende que la entropía de cada región converge a la del universo y que la distribución de la E_m tiene valor esperado de 0. De aquí se sigue que la validación de la hipótesis AEC básicamente prueba la hipótesis nula $\mathcal{H}_0 : \mu_e = 0$, donde μ_e es la media de la distribución de la variable aleatoria E_m.

7.3.1. Prueba de regionalización eficiente

El establecimiento de regiones basado en la divergencia ϕ como criterio de construcción MST y la remoción de aristas mediante un mecanismo híbrido, lleva en forma directa a demostrar que el corolario de regionalización óptima donde se contrasta contra una administrativa se satisface.

En forma adicional a la demostración por construcción del corolario, la calidad de la regionalización se puede determinar al obtener la distribución empírica de la entropía media a través de permutaciones en las unidades básicas de estudio. Para tal efecto se denominará R_ϕ a la regionalización obtenida mediante el criterio de mínima divergencia ϕ.

Dada R_ϕ cuya entropía media es E_ϕ, mediante permutaciones de las unidades básicas entre las unidades agrupadas del universo de estudio, se calcula la entropía media E_a y

de esta se deriva la distribución empírica F_n^a. Una vez establecida F_n^a se obtiene el cuantil y el valor de p asociado al valor de contraste E_ϕ.

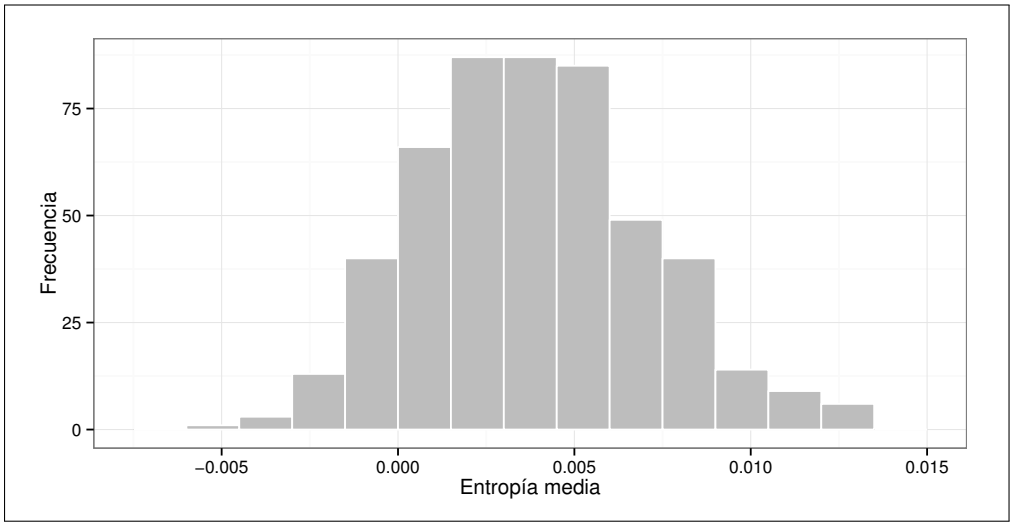


Figura 7.1: *Distribución empírica de la entropía media*

El valor p relativo a la distribución F_n^a de la figura 7.1 es de 0.0052. Por tanto, dada la hipótesis

$$\mathcal{H}_0 : E_\phi = E_a$$
$$\mathcal{H}_1 : E_\phi > E_a$$

hay evidencia estadística suficiente para rechazar $\mathcal{H}_0 : E_\phi = E_a$ con un nivel de confianza estimado de $1 - \alpha = 0.9948$.

7.4. Regionalización con ingreso

Se compara la regionalización obtenida utilizando solo la divergencia ϕ contra la divergencia ponderada $\phi^* = \phi * w$, donde w se calcula por la expresión (5.21) dada en la definición 16 tomando la función $g(x) = e^x$ como ponderador. La partición ponderada

obtenida para una resolución de 6 regiones mostrada en la figura 6.8 no difiere de la que se obtiene sin aplicar una penalización. Este resultado es razonable puesto que que la distribución del ingreso se realiza para un conjunto de intervalos único, lo que en forma natural establece ya una ponderación natural. Un efecto distinto se observaría si en cada municipio se segmenta en forma relativa a este el ingreso.

Una vez establecida la regionalización mediante divergencia ϕ en Coahuila, esta se contrasta con las regiones administrativas.

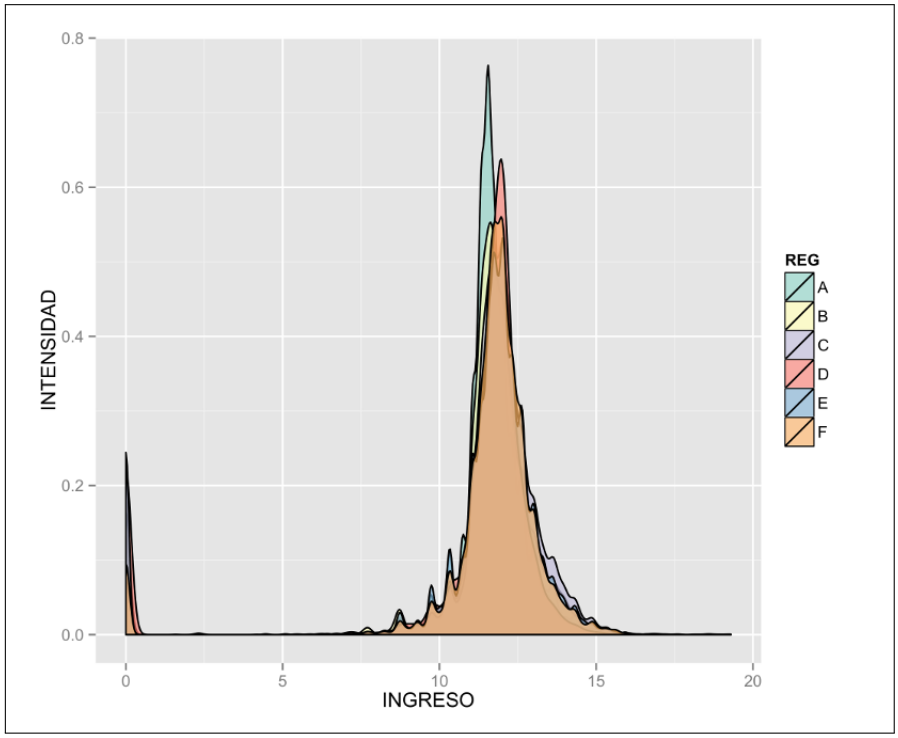


Figura 7.2: *Kernel para regiones de ingreso*

El kernel de las 6 regiones se presenta en la figura 7.2. La entropía por región se muestra en la figura 7.3. La entropía media para regiones administrativas es 0.01036 mientras que la de divergencia es 0.0323. La obtenida mediante la divergencia ϕ es tres veces mayor que la administrativa. Una posible manera de establecer la magnitud de esta diferencia es aplicando bootstrap para construir intervalos de confianza empíricos.

Se debe tomar en cuenta que la mínima divergencia establece la región óptima en la variable de estudio, sin embargo, el contexto y la riqueza de los datos fuente es

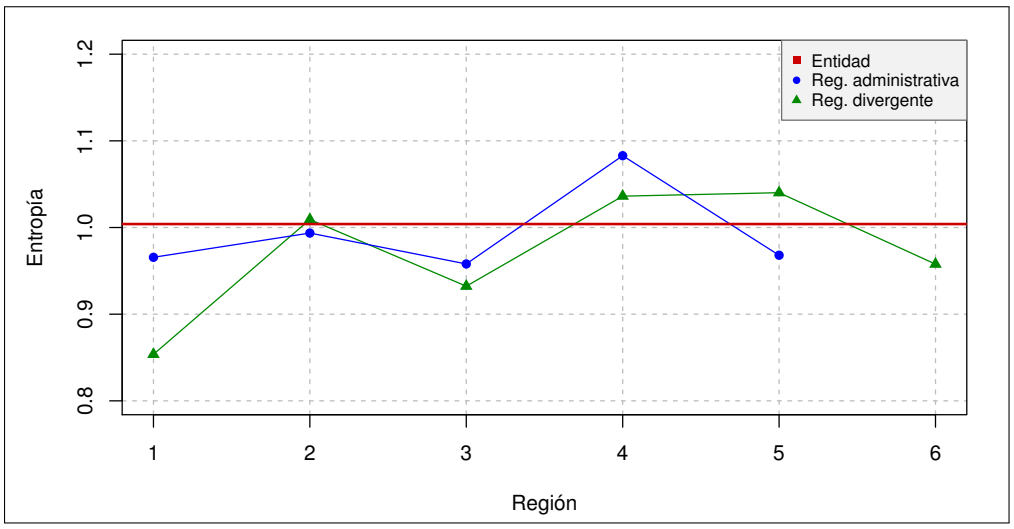


Figura 7.3: *Entropía para regiones de ingreso*

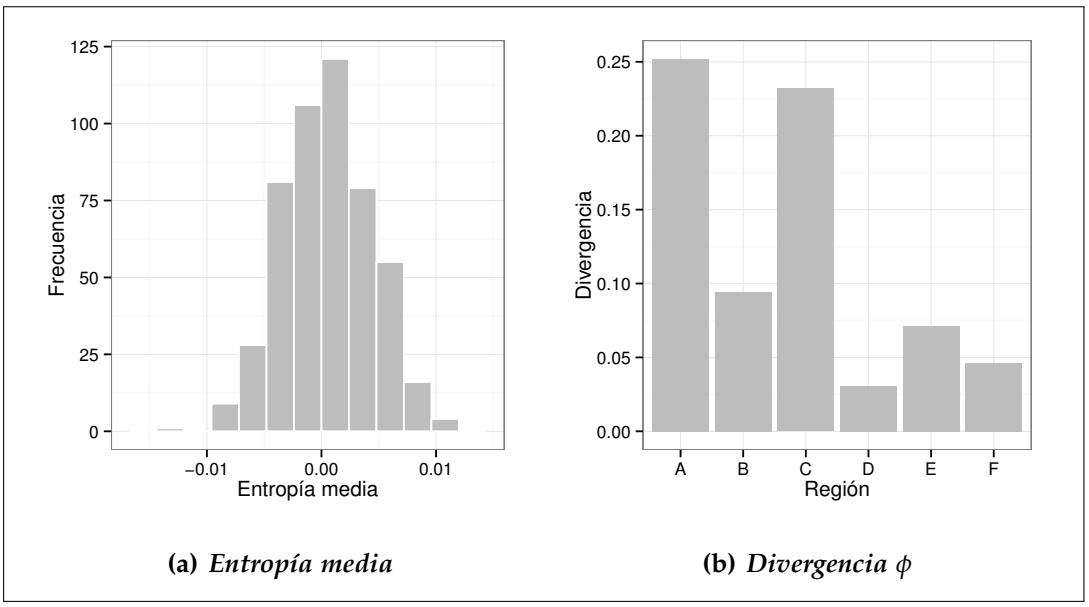
fundamental en su interpretación. Se debe prestar especial atención a la manera en que se realiza la encuesta de ingresos y el aspecto que mide. En este caso es el ingreso por trabajo a nivel de individuo.

La entropía media de la regionalización es $E_\phi = 0.032$. El valor de p asociado a la hipótesis

$$\mathcal{H}_0 : E_\phi = E_a$$
$$\mathcal{H}_1 : E_\phi > E_a$$

es de $p = 0.00062$. Por tanto con una confiabilidad de $1 - \alpha = 0.99938$ se rechaza la hipótesis $\mathcal{H}_0$. Por tanto, se concluye que la partición obtenida es una regionalización distinguible de una agrupación aleatoria.

En la figura 7.4(b) se muestra la divergencia ϕ del ingreso para cada una de las regiones detectadas con relación a la distribución de la entidad completa. Esto nos da una perspectiva de cuáles son las regiones que más se asemejan, en la distribución del ingreso, a la observada en todo el estado. Las de menor divergencia son las regiones



(a) ***Entropía media*** **(b)** ***Divergencia*** ϕ

Figura 7.4: *Entropía media y divergencia para ingreso*

etiquetadas como D, E y F. Las regiones E y F cubren la mayor extensión y cantidad de municipios; estas incluyen a zonas urbanas de municipios mayor población como los son Saltillo (725,123 hab.), Torreón (639,629 hab.) y Monclova (216,206 hab.), lo que representa alrededor del 58 % de la estatal.

La interpretación debe tomarse con reserva, ya que la variable utilizada finalmente es el logaritmo de salarios para incrementar la sensibilidad de la divergencia ϕ.

7.5. Regionalización con el índice de marginación

A partir de los índices de marginación se definen dos particiones (regionalizaciones) de resolución 5 de municipios contigüos: la administrativa y la de mínima divergencia obtenida a partir del método propuesto. La entropía media (ecuación 7.1) para la administrativa es de 0.0470 y la de mínima divergencia de 0.2398. La administrativa tiende a parecerse más a una selección aleatoria en referencia a la variable de estudio.

La figura 7.5 muestra que 3 de las regiones coinciden en entropía con las de divergencia, una difiere en menor medida y otra difiere en mayor magnitud. En la figura 7.6 se

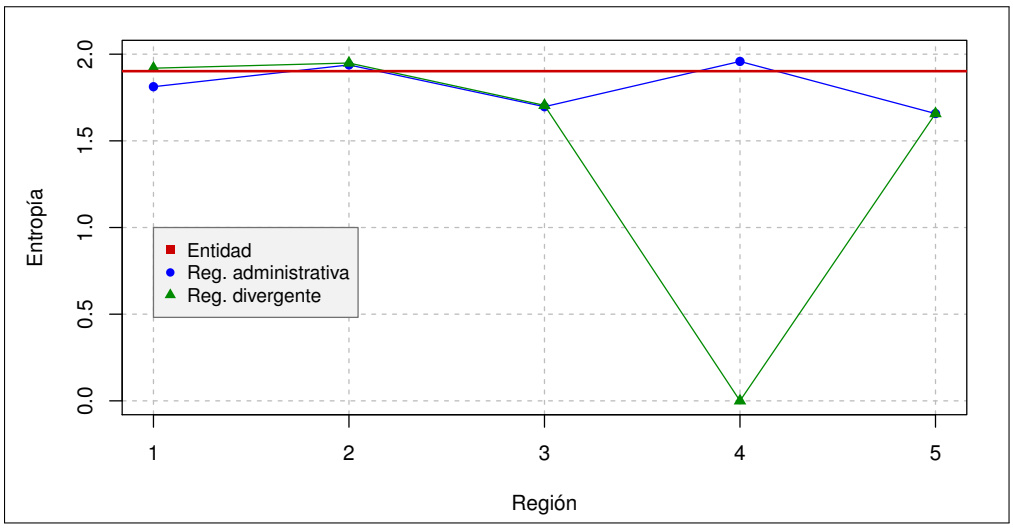


Figura 7.5: *Entropía para regiones de marginación*

muestran los *kernel* obtenidos a partir de la distribución empírica para cada una de las regiones de mínima divergencia.

Los kernel con simetría a la izquierda representan a las regiones de más bajo índice de marginación. Conforme la simetría se recorre a la derecha se incrementa el valor de este índice. La región identificada con más baja marginación coincide con la administrativa que corresponde a la Laguna.

En la regionalización basada en el índice de marginación rural se observa que las regiones detectadas tienden a tener como centro zonas urbanas. El caso más distintivo, y el que muestra menor índice de marginación, es el conjunto que coincide con la región administrativa Laguna de Coahuila. En este grupo se observa una concentración de localidades rurales muy cercanas a la zona conurbada de Torreón.

En contraste con la zona de menos marginación, la de mayor corresponde al conjunto de municipios que va desde Parras al sur del estado hasta Acuña al norte. Se incluyen en esta región los municipios de Cuatro Ciénegas, Sierra Mojada y Ocampo, municipios con menos zonas urbanas y de menor tamaño, además de presentar una gran dispersión geográfica. Esta zona cubre la mayor parte de la región centro-desierto de la entidad.

La región detectada de menor tamaño es formada por los municipios de Castaños y

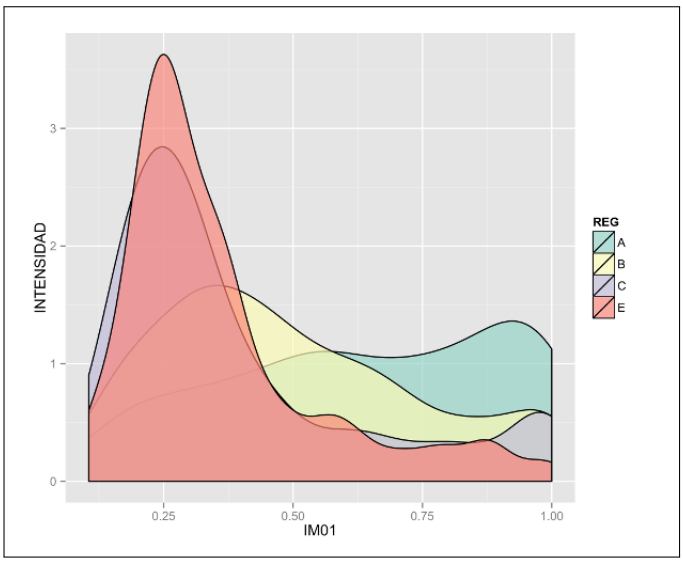


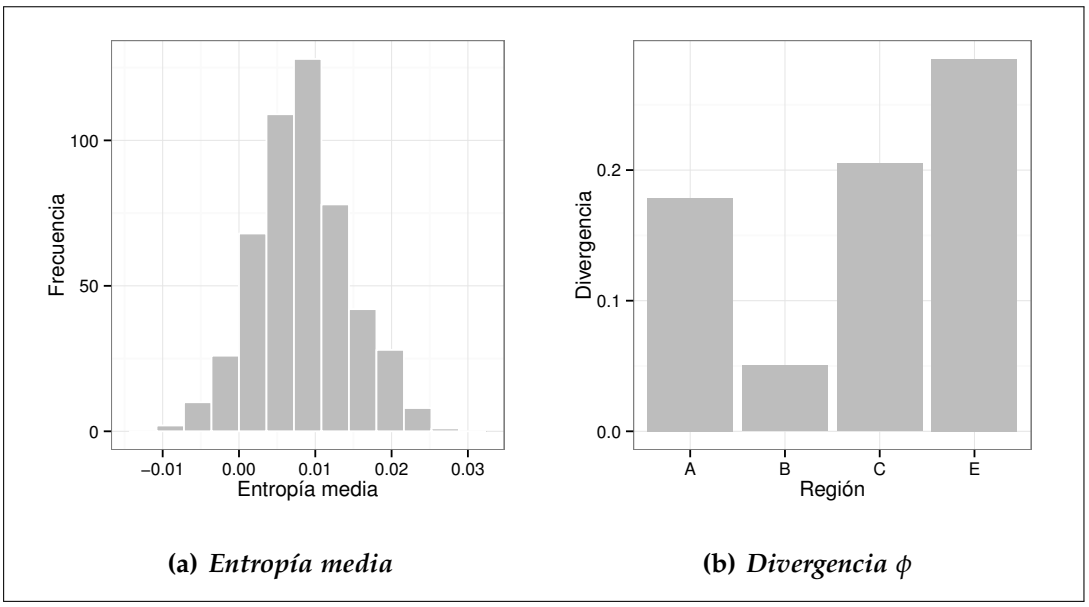
Figura 7.6: *Kernel para regiones de marginación*

Lamadrid con población de 2,314 y 1,835 habitantes respectivamente. Además son los dos municipios que muestran menos dispersión en la distribución empírica del índice de marginación rural.

Se obtiene la distribución empírica de la entropía media para las regiones realizando permutaciones de los índices de marginación. Para tal efecto, se consideran las cuatro regiones detectadas ya que en los municipios más pequeños solo se cuenta con datos de una localidad. Como criterio heurístico, estas pueden integrarse a cualquiera de las dos regiones colindantes.

Una vez establecida una partición se hace bootstrap con 500 muestras en cada región, se obtiene la entropía media con la distribición de todo el estado. Se repite este proceso 500 veces para tener finalmente la distribución empírica de la figura 7.7(a).

La entropía media de la regionalización es $E_\phi = 0.051$. El valor de p asociado a la hipótesis



(a) ***Entropía media*** **(b)** ***Divergencia*** ϕ

Figura 7.7: *Entropía media y divergencia para IM01*

$$\mathcal{H}_0 : E_\phi = E_a$$
$$\mathcal{H}_1 : E_\phi > E_a$$

es de $p = 0.0085$. Por tanto con una confiabilidad de $1 - \alpha = 0.9915$ se rechaza la hipótesis $\mathcal{H}_0$. Por tanto, se concluye que la partición obtenida es una regionalización distinguible de una agrupación aleatoria.

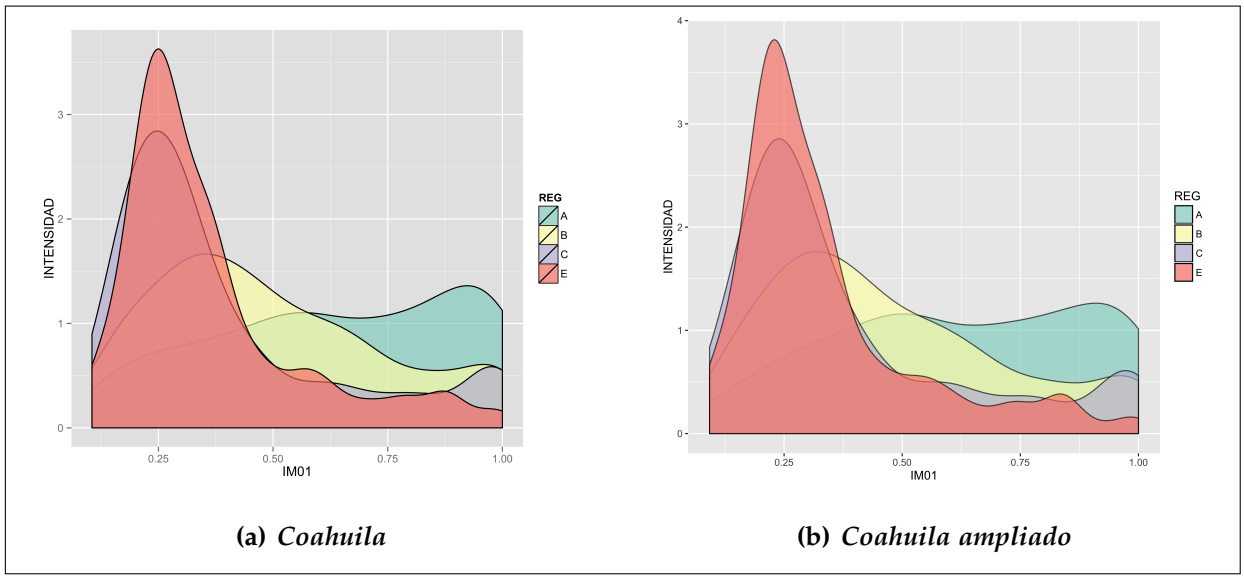
En la figura 7.7(b) se muestra la divergencia ϕ del *IM01* para cada una de las regiones detectadas con relación a la distribución de la entidad completa. Esto nos da una perspectiva de cuál es la región que más se asemeja, en la intensidad de marginación, a la observada en todo el estado. La de menor divergencia es la región etiquetada como *B* que corresponde a la similar a la región administrativa sur (figuras 6.11 y 7.6). En esta región se ubica la la capital del estado Saltillo en zona conurbada con Ramos Arizpe.

7.5.1. Comparación con Coahuila ampliado del índice de marginación

Se toma como referencia de análisis la figura 7.8 donde se comparan los kernels del *IM01* rural en Coahuila (a) y Coahuila ampliado (b).

La similitud de los kernels se debe, en términos generales, a que el comportamiento de las localidades rurales en lo que al *IM01* se refiere, es similar en los municipios colindantes de cada una de las regiones establecidas con las localidades de Coahuila.

El caso más significativo es el de la región que coincide con la región administrativa de la Laguna. El nodo de referencia es la zona conurbada que forma parte de la zona metropolitana Laguna. Esta zona la conforman básicamente los municipios de Torreón en Coahuila y Lerdo en Durango. Los municipios de Tlahualilo y General Simón Bolívar de Durango, con pocas localidades rurales, presentan valores similares de *IM01* a los municipios de Francisco y Madero, Torreón (en su parte sur) y Viesca en Coahuila.



(a) ***Coahuila*** **(b)** ***Coahuila ampliado***

Figura 7.8: *Kernel de marginación rural para regiones de Coahuila (análisis)*

Contrario a la región centrada en la Laguna, la parte de más alta marginación de Coahuila la conforman los municipios de Ocampo, Sierra Mojada, Cuatro Ciénegas y Parras. Los municipios colindantes de entidades vecinas son Jiménez, Camargo y Manuel Benavides en Chihuahua, además de Melchor Ocampo en Zacatecas. Los muncipios

colindantes con Chihuahua comparten la condición de zona desértica y de alta dispersión en sus localidades. En la parte de Parras, al sur de Coahuila, aunque hay menos dispersión al igual que en Melchor Ocampo de Zacatecas, la distribución empírica del *IM01* es similar en ambos casos.

La región identificada en la parte sureste de Coahuila colinda con los municipios de Mina, García, Santa Catarina, Santiago y Galeana en Nuevo León, además de El Salvador y Concepción del Oro en Zacatecas. Estos municipios colindan con Saltillo, Arteaga, Ramos Arizpe y Castaños. Dado que la mayoría de las localidades rurales en estos municipios externos están cercana a localidades de Coahuila pero alejados de la zona urbana, presentan similares niveles de marginación.

En la parte noreste de Coahuila, los municipios de Hidalgo, Juárez, Progreso y Candela colindan con los municipios de Anahuac, Lampazos de Naranjo y Bustamante en Nuevo León. En los municipios de ambos estados se observa una alta dispersión de localidades rurales, situación que preserva la distribución del índice de marginación en Coahuila al considerar los municipios colindantes.

Las regiones obtenidas tienen concordancia con las zonas metropolitanas establecidas a nivel nacional. En Coahuila, conjuntamente con Durango se conforma la zona Laguna. Las zonas metropolitanas con municipios ancla Saltillo y Monclova generan, en forma conjunta otra de las regiones de marginación. La zona con municipio ancla Piedras Negras está contenida en la segunda zona en magnitud establecida para marginación. La zona de marginación de mayor extensión no corresponde a alguna zona metropolitana, siendo esta la de mayor dispersión de localidades rurales y más alta marginación.

7.6. Anotaciones

De los resultados obtenidos se observa que la regionalización obtenida para marginación tiende a darle más peso a las concentraciones urbanas. La región de mayor extensión geográfica y mayor dispersión de localidades urbanas presenta la más alta intensidad de marginación. En contraste para el ingreso, esto no ocurre en la misma medida ya

que, particularmente, una de las regiones conecta al municipio de Torreón con el de Sierra Mojada que es uno de los que presentan mayor dispersión y menor densidad de población[1].

[1]La densidad de población se considera el cociente del número de habitantes entre el área del municipio.

CAPÍTULO 8

CONCLUSIONES

La metodología propuesta de regionalización basada en un criterio de mínima divergencia no pretende rivalizar con una regionalización administrativa. El objetivo es más bien aportar elementos que sirvan para dos propósitos: el primero es revelar, de una manera más precisa, características y correlaciones entre factores de estudio; el segundo es utilizar los resultados como una herramienta de análisis que provea información relevante como soporte en el establecimiento o adecuación de políticas públicas.

Se enfatiza en un visión reduccionista en contraste a una holista cuando se trata de incorporar características medibles, esto es, ver cada unidad como una colección de partes que la conforman y no como un valor agregado. Esto revela información granular al momento de efectuar comparaciones o se pretenda clasificar en grupos las unidades de estudio.

¿Cuál es la relevancia de considerar la distribución empírica de las características socio-económicas de una unidad espacial como criterio de regionalización?

Si bien esta no es la primera pregunta que se plantea en esta investigación, es pertinente darle primero respuesta ya que en torno a este elemento se construye la metodología propuesta para delimitar regiones eficientes en el sentido de la información que aportan.

La importancia de utilizar un enfoque de información que incorpora la distribución em-

pírica como argumento para establecer disimilitudes, en contraste con valores agregados, radica en que en este se enmascara información relevante al tipificar el objeto de estudio. El sentido estocástico de la propuesta difiere de los modelos o métricas deterministas al establecer distancias, de corte cuantitativo, entre unidades de estudio.

El uso de la distribución de probabilidad es la contraparte matemática del enfoque reduccionista, considerado no en un sentido mecanisista de ver a una estructura de nivel relativamente alto como la suma de sus partes de más bajo nivel. Se entiende más bien en el sentido emergista, donde dicha estructura de alto nivel no solo es la suma de las partes sino que también se deriva de interacciones implícitas entre estas. La distribución de probabilidad y la concepción reduccionista son las expresiones equiparables al espacio matemático y el espacio económico.

¿Qué tan alejados están, en términos de sus características, los objetos espaciales contigüos geográficamente?

Esta pregunta tiene diversas respuestas dependiendo en el contexto en que se encuadre y enfoque a utilizar. Existe una variedad de métricas utilizadas para establecer distancias y disimilitudes entre objetos de estudio. Se pueden utilizar criterios básicos que van desde comparar considerando a la población o la extensión territorial, pasando por métricas comunes como distancias euclideanas y familias de distancias no euclideanas, hasta medidas más complejas que incorporen factores y sus interacciones.

La distancia en que se encuentran dos unidades de estudio se obtiene desde la perspectiva de información útil que deriva en un enfoque evolutivo: las regiones cambian de acuerdo al contexto, en el tiempo y en su ubicación geográfica. La medida de lejanía, disimilitud o divergencia está dada en términos de la información que aportan dos unidades de referencia que se comparan.

Una de las características a destacar en una clasificación regional mediante una medida de divergencia es que en principio no pondera sobre la población de estudio. Esto significa que no discrimina por densidad, sino clasifica de acuerdo a la intensidad del fenómeno de estudio.

Un caso claro donde no se debe ponderar por densidad poblacional es el caso de marginación. Esto se debe a que independientemente de la cantidad de población, dos distribuciones de probabilidad del índice reflejan estructuras similares, es decir, no por ser menor una población se está menos marginado aunque la estructura estocástica de su índice comparada con poblaciones mayores sea similar.

Desde una perspectiva de inclusión social es fundamental dimensionar el problema de estudio por su significado como fenómeno y no por la cantidad de población en que incide. La ponderación por densidad poblacional estaría enfocada a la planeación de estrategias para la aplicación o diseño de políticas públicas.

Si bien para el caso de marginación una ponderación por la densidad dada en la definición 16 no se utiliza, en otros contextos es aplicable. Aunque en la distribución del ingreso va implícita una ponderación al utilizar una categorización común de clasificación para todos los municipios, la aplicación del ponderador enfatiza esta diferencia, de ahí la recomendación de su integración como un elemento para incrementar la sensibilidad del método.

¿En qué medida se distingue un objeto espacial (o conjunto) de todo el sistema (conjunto de objetos) en el que se encuentra inmerso?

El uso de árboles generados mínimos utilizando la medida de divergencia ϕ mostró ser eficiente en detectar las unidades geográficas contiguas considerando la distribución de probabilidad empírica de las variables que las tipifican. La métrica fue lo suficientemente sensible para detectar mínimas diferencias entre las distribuciones empíricas generadas mediante simulación.

Dado que la métrica ϕ es capaz de detectar diferencias inducidas mediante simulación, la capacidad de distinguir distintos grupos se traslada al terreno empírico utilizando datos reales.

Si bien la metodología propuesta se basa en la existencia de datos suficientes para establecer regiones, esta también sirve como elemento para el diseño de encuestas y toma de datos enfocada a delimitar estratos ad-hoc en función de factores de interés

previamente establecidos. El desarrollo de tecnologías para toma de datos o análisis basado en minería de datos potencia la aplicación del método en la definición de regiones alto grado de desagregación.

La partición del universo de estudio obtenida a partir de la homogeneidad interna de los grupos y la heterogeneidad entre grupos es una formulación de lo que se definió como región homogénea. Este modelo de regionalizar se puede extender y enriquecer sustancialmente si se tiene información relativa a las interacciones entre los agentes que las conforman. Esto es, encuadrarlo en la visión emergista de considerar tanto las partes de un sistema como los efectos de sus interacciones.

¿En qué magnitud difiere una regionalización de otra tomando en cuenta características cuantitativas que las definen?

De igual manera que la pregunta relativa a establecer una métrica de divergencia, la comparación entre diversas regionalizaciones es una discusión latente. Aquí se centra el estudio en establecer una función objetivo a optimizar. Si bien en la literatura se hace un planteamiento general de la estructura de esta función, en muchos de los casos se centra en comparar la variablidad de las características interregionales.

Aquí se establece, para ser consistentes con el enfoque de información, a la entropía media como mecanismo de diferenciación entre regionalizaciones. Esta medida reveló las diferencias en el sentido de qué tanto se aleja una regionalización dada a una obtenida mediante una selección aleatoria. Dado que se parte de una inducción apriori in silico de regiones para validar su capacidad de distinguirlas, la prueba de la hipótesis *AEC* es válida en otros contextos, particularmente en las aplicaciones con datos reales.

Para el caso de marginación, calculado como colección de la marginación en localidades rurales, la regionalización obtenida revela que:

- Las localidades rurales con menos marginación tienden a aglutinarse en torno a las zonas urbanas. Esto muestra, en una primera instancia, que la proximidad a estas concentraciones favorece en la reducción de las carencias asociadas a este índice.

- Uno de los factores que asocia a municipios contiguos en una misma región es el grado de dispersión de sus localidades rurales. Aunque esta variable está implícita en el cálculo del *IM01* a nivel municipal, por la baja sensibilidad del índice a la remoción de alguno de sus factores no muestra la importancia que tiene, sin embargo toma relevancia una vez que se agrupan tomando como base su divergencia.

- La distribución empírica del *IM01* de localidades rurales en la región identificada como *D* es la que más similitud presenta a la empírica de toda la entidad. Esta región coincide en su mayoría con la región administrativa sureste, agregando los municipios de Abasolo, Castaños y Frontera, los cuales tienen pocas localidades en comparación a Saltillo, Ramos Arizpe y Arteaga.

- La inclusión de municipios colindantes de entidades vecinas a Coahuila no modifica sustancialmente la distribución empírica del *IM01* calculado solo para Coahuila. Esto es razonable dado que las entidades no son conjuntos aislados y las condiciones permean entre localidades rurales cercanas. Finalmente, lo que ocurre es que lo observado en las localidades rurales de Coahuila cercanas a las de las entidades vecinas es un reflejo de estas, al menos en esta variable de estudio.

- Las regiones de marginación obtenidas mediante divergencia de información muestran concordancia con las zonas metropolitanas establecidas a nivel nacional. Esto se había ya observado en relación a la concentración de localidades rurales de baja marginación cercanas a zonas conurbadas.

- El hecho de que la inclusión de localidades rurales de municipios colindantes en entidades vecinas a Coahuila no cambie sustancialmente la distribución empírica del *IM01*, no garantiza que en un caso general al regionalizar todo el país eso ocurra en otras entidades vecinas. Las condiciones particulares de Coahuila, Chihuahua, Durango, Nuevo León y Zacatecas en sus municipios colindantes arrojaron el resultado mostrado en el capítulo de análisis.

En relación a la regionalización obtenida para el ingreso, se destacan tres aspectos principales:

- Las regiones de mayor extensión concentran a la mayoría de la población al incluir las zonas conurbadas de Saltillo, Torreón y Monclova;
- las zonas conurbadas representan la fuente de trabajo de localidades en municipios aledaños a estas. Esto es, la movilidad laboral contribuye a aglutinar municipios con baja densidad poblacional;
- el ingreso es calculado a partir de individuos y no del lugar donde está ubicada la fuente de empleo. Esto conlleva trasladar esa característica al lugar de residencia de la persona, que en muchos de los casos se mueve de una zona rural a una zona urbana para laborar.

Como contraste entre la regionalización por ingreso y la obtenida con el índice de marginación se destaca una diferencia sustancial: mientras que la marginación es geográficamente estática, el ingreso no necesariamente tiene su fuente en el municipio de residencia. Particularmente las variables de marginación ponderan las carencias nivel vivienda y el ingreso como individuo.

Otro factor que hace la diferencia en las regiones *IM01* e ingreso, es el hecho de que la unidad de estudio en marginación es la localidad y en ingreso es el individuo.

Finalmente, en lo referente al método general de regionalización propuesto, se destacan dos aspectos centrales:

1. El método propuesto es generalizable al caso multivariado, donde además se puede extender la función de ponderación, similar a un modelo gravitatorio, para establecer interacciones entre las unidades agrupadas de estudio;
2. es fundamental considerar que no existe un enfoque único de lo que es una región, y que cada una de estos enfoques no necesariamente aporta, de manera directa, un método específico para tipificar una región. Esto significa que la mecánica a seguir para regionalizar un universo de estudio puede tener diversas rutas y resultados, particularmente en los sistemas complejos, dentro de los cuales se enmarcaría una economía.

BIBLIOGRAFÍA

Adler, J. (2009). *R IN A NUTSHELL*. O'Reilly.

Assuncão, R. M., Neves, M. C., Camara, G., y Da Costa Freitas, C. (2006). Efficient regionalization techniques for socio–economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20(7):797–811.

Ayres, R. U. (1994). *Information, Entropy, and Progress*. AIP Press.

Ball, P. (2010). *Masa crítica, cambio, caos y complejidad*. Fondo de Cultura Económica.

Batty, M. (1978). *Speculations on an information theoric approach to spatial representation*, chapter Spatial representation and spatial interaction: an overview, pages 115–147. Studies in Applied Regional Science.

BiografiasyVidas (2000). William stanley jevons. `http://www.test.org/doe/`.

Bivand, R. S., Pebesma, E. J., y Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with R*. Springer.

Brenner, N. (2000). Building "euro-regions": Locational politics and the political geography of neoliberalism in post-unification germany. *European Urban and Regional Studies*, 7(4):319–345.

Bryant, J. (2012). *Thermoeconomics. A Thermodynamic Approach of Economics*. VOCAT.

CONAPO (2010a). Anexo c: Metodología de estimación del índice de marginación. Technical report, Consejo Nacional de Población.

CONAPO (2010b). índice de marginación por entidad federativa y municipio 2010. Technical report, Consejo Nacional de Población.

CONAPO (2013). índice absoluto de marginación 2000-2010. Technical report, Consejo Nacional de Población.

Cooper, R., Donaghy, K., y Hewings, G. (2010). Preface. *Globalization and Regional Economic Modeling*, 1:V.

Cortés, F. y Rubalcava, R. M. (1984). *Técnicas estadísticas para el estudio de la desigualdad social*. El Colegio de México.

Cover, M. T. y Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley & Suns.

Diller, A. (1999). *LATEX, Line by Line*. Wiley.

Donaghy, K. P. (2010). Globalization and regional economic modeling: Analytical and methodological chalenges. *Globalization and Regional Economic Modeling*, 1:1–11.

DuBois, P. (2006). *MySQL Cookbook*. O'Reilly.

Eaton, J. W., Bateman, D., y Hauberg, S. (2008). *GNU Octave Manual*.

Farrell, H. y Héritier, A. (2005). A rationalist-institutionalist explanation of endogenous regional integration1. *Journal of European Public Policy*, 12(2):273–290.

Floridi, L. (2010). *Information, A Very Short Introduction*. OXFORD University Press.

Fotheringham, A. S., Brunstdon, C., y Charlton, M. (2000). *Quantitative Geography*. SAGE.

Fotheringham, A. S. y Wong, D. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, 23(7):1025–1044.

Fundation, T. D. (2015). `https://es.libreoffice.org/`.

Gasca, Z. J. (2009). *Geografía regional. La región, la regionalización y el desarrollo regional en México*. Instituto de Geografía UNAM.

Georgescu-Roegen, N. (1999). *The Entropy Law and the Economic Process*. Harvard.

Goodwin, M. (2012). *Economix*. ABRAMS.

Grandrud, C. (2015). *Reproducible Research with R and RStudio*. Chapman and Hall/CRC.

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, 78(6):1360–1380.

Hey, T., Tansley, S., y Tolle, K. (2009). *The Fourth Paradigm: Data Intensive Scientific Discovery*. Microsoft Research.

Hofstadter, D. R. (1982). *Gödel, Escher, Bach: una eterna trenza dorada*. Consejo Nacional de Ciencia y Tecnología.

Hofstadter, D. R. (1999). *Gödel, Escher, Bach: an Eternal Golden Braid*. Basic Books.

INEGI (2011). Síntesis metodológica y conceptual del censo de población y vivienda 2010. Technical report, Instituto Nacional de Geografía e Historia.

Jan, H. R. v. D. y Gijsbertus, P. v. W. (2010). Globalization and intermodal transportation: Modeling terminal locations using a three-spatial scales framework. *Globalization and Regional Economic Modeling*, 1:133–152.

Janssens, J. (2015). *Data Science at the Command Line*. O'Reilly.

Jaynes, E. T. (1991). How we should use entropy in economics. Technical report, St. John's College, Cambridge.

Jobson, J. D. (1991). *Applied Multivariate Data Analysis. Volume II: Categorical and Multivariate Methods*, volume II. Springer Verlag.

Konishi, S. y Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. Springer Series in Statistics.

Krugman, P. (1991). Increasing returns and economic geography. *Journal of Politic Economy*, 99:483–499.

Kummel, R. (2011). *The Second Law of Economics*. Springer.

Lu, M. (2011). Ad hoc regionalism in rural development. *The Geographical Review*, 3:334–352.

Lüchinger, R. (2007). *Los 12 economistas más importantes de la historia*. Grupo Editorial Norma.

Masser, I. y Brown, P. (1978a). *Spatial represantation and spatial interaction*, chapter Spatial representation and spatial interaction: an overview, pages 1–23. Studies in Applied Regional Science.

Masser, I. y Brown, P. (1978b). *Spatial representation and spatial interaction*. Studies in Applied Regional Science.

Mihalcea, R. y Radev, D. (2011). *Graph-Based Natural Language Processing and Information Retrieval*. Cambridge University Press.

Milton, J. S. y Arnold, J. C. (1990). *Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computing Sciences*. Mc Graw Hill.

Mitchell, M. (2009). *Complexity, A Guided Tour*. OXFORD University Press, kindle edition.

Mumford, S. y Lill, R. A. (2013). *Causation*. Oxford University Press.

Openshaw, S. (1984). The modifiable aereal unit problem. *Concepts and Techniques in Modern Geography*, 38:1–40.

O'Sullivan, D. y Unwin, D. J. (2010). *Geographic Information Analysis*. Wiley.

Paasi, A. (2004). Place and region: looking through the prism of scale. *Progress in Human Geography*, 28(4):536–546.

Quarteroni, A., Saleri, F., y Gervasio, P. (2010). *Scientific Computing with MATLAB and Octave*. Springer.

Rionda R., J. I. (2005). *Contextos del desarrollo regional en México*. eumed.net.

Rosser, B. (2011). *Complex Evolutionay Dynamics in Urban-Regional and Ecologic-Economic Systems*. Springer.

Roy, J. R. y Thill, J. C. (2004). Spatial interaction models. *Regional Science*, 83:339–361.

Rudy, A. P. (2005). Imperial contradictions: is the valley a watershed, region, or ciborg? *Journal of Rural Studies*, 21:19–38.

Saslow, W. M. (1999). An economic analogy to thermodynamics. *American Journal of Physics*, 67:1239–1247.

Sherman, G. (2012). *The Geospatial Desktop*. Local-e Press.

Storper, M. (1997). *The Regional World*. Guilford.

Taverna (2015). `http://www.taverna.org.uk/introduction/what-is-in-silico-experimentation/`.

Vickerman, R. (2007). Transport, globalization and the changing concept of the regional. *Globalization and Regional Economic Modeling*, 1:35–43.

Waller, L. A. (2009). *The SAGE Handbook of Spatial Analysis*, chapter Detection of Clustering in Spatial Data, pages 299–320. SAGE.

Wehenkel, L. (2003). Théorie de l'information et du codage. Technical report, Université de Liége, Faculté de sciences appliqués.

Wing, I. S. y Anderson, P. W. (2010). Modeling small area economic change in conjuntion with a multiregional cge model. *Globalization and Regional Economic Modeling*, 1:263–288.

Wixted, B. y Cooper, J. R. (2010). The evolution of oecd ict inter-cluster networks 1970-2000: An input-output study on the interdependencies between nine oecd economies. *Globalization and Regional Economic Modeling*, 1:153–182.

APÉNDICE A

CONCEPTOS Y MODELOS MATEMÁTICOS

La construcción de clusters está directamente relacionada con la métrica que se utilice las cuales determinan distancias entre objetos espaciales.

Definición 18 *Sea* $\mathbf{X} \in \Re^{n \times a}$ *la matriz representada como un conjunto de renglones:*

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^t \\ \mathbf{x}_2^t \\ \vdots \\ \mathbf{x}_n^t \end{pmatrix}$$

donde $\mathbf{x}_i \in \Re^a$, $i = 1, 2, \ldots, n$. *Se definen las siguientes distancias entre los vectores* $\mathbf{x}_r$ *y* $\mathbf{x}_s$ *(Jobson, 1991, p 487-508):*

- *Euclideana*

$$\begin{aligned} d_{rs}^2 &= \sum_{j=1}^{a} (x_{rj} - x_{sj})^2 \\ &= ||\mathbf{x}_r - \mathbf{x}_s||^2 \end{aligned}$$

- *Euclidena ponderada*

$$\begin{aligned} d_{rs}^2 &= \sum_{j=1}^{a} w_j (x_{rj} - x_{sj})^2 \\ &= (\mathbf{x}_r - \mathbf{x}_s)^t \mathbf{D}^{-1} (\mathbf{x}_r - \mathbf{x}_s) \\ &= ||\mathbf{x}_r - \mathbf{x}_s||_w^2 \end{aligned}$$

  donde $\mathbf{D} = diag(w_1, w_2, \ldots, w_a)$.

- *Euclidena con centroide*

$$\begin{aligned} d_{rs}^2 &= 2 \left[\sum_{j=1}^{a} (x_{rj} - \bar{x}_{\cdot j})^2 + \sum_{j=1}^{a} (x_{sj} - \bar{x}_{\cdot j})^2 \right] \\ &= 2 \left(||\tilde{\mathbf{x}}_r||^2 - ||\tilde{\mathbf{x}}_s||^2 \right) \end{aligned}$$

  donde $\bar{x}_{\cdot j} = \frac{1}{2}(x_{rj} + x_{sj})$ *y* $\tilde{\mathbf{x}}_r = \mathbf{x}_r - \bar{\mathbf{x}}_{rs}$.

- *Mahalanobis*

$$d_{rs}^2 = (\mathbf{x}_r - \mathbf{x}_s)^t \mathbf{S}^{-1} (\mathbf{x}_r - \mathbf{x}_s)$$

  donde $\mathbf{S} = Var(\mathbf{X})$.

- *Manhattan*

$$d_{rs} = \sum_{j=1}^{a} |x_{rj} - x_{sj}|$$

- *Minkowsky*

$$d_{rs} = \left[\sum_{j=1}^{a} |x_{rj} - x_{sj}|^{\lambda} \right]^{\frac{1}{\lambda}}$$

- *Canberra*

$$d_{rs} = \sum_{j=1}^{a} \frac{|x_{rj} - x_{sj}|}{|x_{rj} + x_{sj}|}$$

- *Máximo (norma suprema)*

$$d_{rs} = max|x_{rj} - x_{sj}|$$

□

La distancia de *Minkowsky* con $\lambda = 1$ corresponde a la de *Manhattan* y con $\lambda = 2$ a la *Euclideana*. Más aún, esta distancia satisface que $d_{rs} = 0$ solo si $\mathbf{x}_r = \mathbf{x}_r$ y $d_{rs} \leq d_{rm} + d_{ms}$ $\forall r, s, m$.

Teorema 4 *Sean $g(x)$ y $h(x)$ dos funciones continuas tales que $g(x) < h(x)$*

$$\frac{d}{d\lambda} \int_{t_0}^{g(\lambda)} f(x, \lambda) dx = f(g(\lambda), \lambda) g'(\lambda) + \int_{t_0}^{g(\lambda)} \frac{\partial}{\partial \lambda}(x, \lambda) dx$$

Regla de Leibniz

□

Definición 19 *Sea X una variable aleatoria con al menos los tres primeros momentos finitos. La* asimetría, *respecto a la media, de X, denotada como γ, está dada por la expresión*

$$\gamma = \frac{E[X - E(X)]^3}{\sigma^3}$$

donde σ es la desviación estándar de X.

□

Teorema 5 *Sea X una variable aleatoria con distribución binomial con parámetros n y p. Si $n \to \infty$ y $p \to 0$ entonces la distribución binomial converge a una distribución Poisson con parámetro $\lambda = np$.*

Demostración

Lo que se debe demostrar es

$$\lim_{\substack{n\to\infty \\ p\to 0}} \mathcal{B}(n,p) = \mathcal{P}(np)$$

$\boxdot$

APÉNDICE B

TERMODINÁMICA E INFORMACIÓN

B.1. Entropía de Boltzmann

Teorema 6 *La entropía de Boltamann es una cantidad que varía proporcionalmente al logaritmo de probabilidad W, en la cual la frecuencia relativa de la muestra obtenida de un modelo especificado es acorde con la distribución verdadera*

Demostración

Sea n el tamaño de una muestra independiente de la distribución f y suponga que se conocen tanto la distribución de frecuencias $\{n_1, n_2, \ldots, n_k\}$ *como la de frecuencias relativas* $\{g_1, g_2, \ldots, g_k\}$. *Dado que la la probabilidad con la cual la distribución de frecuencias se obtiene es*

$$W = \frac{n!}{n_1! n_2! \cdots n_k!} f_1^{n_1} f_2^{n_2} \cdots f_k^{n_k}$$

Tomando el logaritmo en ambos lados de la ecuación y utilizando la aproximación de Stirling ($\log n! = n \log n - n$) *se sigue:*

$$
\begin{aligned}
\log W &= \log\left(\frac{n!}{n_1!n_2!\cdots n_k!} f_1^{n_1} f_2^{n_2}\cdots f_k^{n_k}\right) \\
&= \log\left(\frac{n!}{n_1!n_2!\cdots n_k!}\right) + \log(f_1^{n_1} f_2^{n_2}\cdots f_k^{n_k}) \\
&= \log n! - \sum_{i=1}^{k} \log n_i! + \sum_{i=1}^{k} n_i \log f_i \\
&\approx n\log n - n - \sum_{i=1}^{k} n_i \log n_i + \sum_{i=1}^{k} n_i + \sum_{i=1}^{k} n_i \log f_i \\
&= n\log n - \sum_{i=1}^{k} n_i \log n_i + \sum_{i=1}^{k} n_i \log f_i \\
&= \sum_{i=1}^{k} n_i \log n - \sum_{i=1}^{k} n_i \log n_i + \sum_{i=1}^{k} n_i \log f_i \\
&= -\sum_{i=1}^{k} n_i (\log n_i - \log n) + \sum_{i=1}^{k} n_i \log f_i \\
&= -\sum_{i=1}^{k} n_i \log\left(\frac{n_i}{n}\right) + \sum_{i=1}^{k} n_i \log f_i \\
&= -\sum_{i=1}^{k} n_i \log g_i + \sum_{i=1}^{k} n_i \log f_i = \sum_{i=1}^{k} n_i (\log f_i + \log g_i) \\
&= \sum_{i=1}^{k} n_i \log\left(\frac{f_i}{g_i}\right) = \frac{n}{n}\sum_{i=1}^{k} n_i \log\left(\frac{f_i}{g_i}\right) \\
&= n\sum_{i=1}^{k} g_i \log\left(\frac{f_i}{g_i}\right) = nB(g;f) \\
&\Rightarrow B(g;f) = \frac{1}{n}\log W
\end{aligned}
$$

□

B.2. Información de Shanon

Definición 20 *Sea X una variable aleatoria discreta con distribución de probabilidad $\{p_1, p_2, \ldots, p_n\}$. La entropía de X, denotada por $H(X)$ se define como*

$$H(X) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

□

Teorema 7 *El valor máximo de la entropía de Shannon $H(X)$ es $\log_2(n)$.*

Demostración

Solo se esboza una parte de la demostración. Dado que el máximo se obtiene cuando la distribución es uniforme, se sigue que:

$$\begin{aligned} H(X) &= -\sum_{i=1}^{n} p_i \log_2(p_i) \\ &= -\sum_{i=1}^{n} \frac{1}{n} \log_2\left(\frac{1}{n}\right) \\ &= -\frac{1}{n}\sum_{i=1}^{n} \log_2(n^{-1}) \\ &= -\frac{1}{n}\sum_{i=1}^{n} -\log_2(n) \\ &= -\frac{1}{n}(-n\log_2(n)) \\ &= \log_2(n) \end{aligned}$$

□

Teorema 8 *Si $F_X \sim U(\mathcal{X})$ entonces*

$$I(g||f) = log_2|\mathcal{X}| - H_g(X)$$

Demostración

Sin pérdida de generalidad consideremos el caso discreto:

$$
\begin{aligned}
I(g||f) &= \sum_{x\in\mathcal{X}} g_x(x)\log_2\frac{g_x(x)}{f_x(x)} \\
&= \sum_{x\in\mathcal{X}} g_x(x)[\log_2 g_x(x) - \log_2 f_x(x)] \\
&= \sum_{x\in\mathcal{X}} g_x(x)\log_2 g_x(x) - \sum_{x\in\mathcal{X}} g_x(x)\log_2 f_x(x) \\
&= -H_g(x) - \log_2\frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} g_x(x) \\
&= -H_g(x) + \log_2|\mathcal{X}| \\
&= \log_2|\mathcal{X}| - H_g(x)
\end{aligned}
$$

$\boxdot$

APÉNDICE C

CUADROS DE REFERENCIA

Cuadro C.1: *Municipios y regiones de Coahuila*

CM	NOMBRE	REGIÓN	POB	ÁREA
05001	Abasolo	CENTRO	1,070	737.96
05002	Acuña	NORTE	136,755	11,461.27
05003	Allende	NORTE	22,675	250.63
05004	Arteaga	SURESTE	225,44	1,619.15
05005	Candela	CENTRO	1,808	2,102.82
05006	Castaños	CENTRO	25,892	3,316.09
05007	Cuatro Ciénegas	CENTRO	13,013	10,588.65
05008	Escobedo	CENTRO	2,901	1,018.16
05009	Francisco I. Madero	LAGUNA	55,676	2,786.11
05010	Frontera	CENTRO	75,215	454.09
05011	General Cepeda	SURESTE	11,682	2,615.30
05012	Guerrero	NORTE	2,091	2,913.40
05013	Hidalgo	NORTE	1,852	1,123.85
05014	Jiménez	NORTE	9,935	2,197.07
05015	Juárez	CARBONIFERA	1,599	2,444.04
05016	Lamadrid	CENTRO	1,835	668.70
05017	Matamoros	LAGUNA	107,160	798.27

05018	Monclova	CENTRO	216,206	1,242.17
05019	Morelos	NORTE	8,207	636.66
05020	Múzquiz	CARBONIFERA	66,834	8,258.30
05021	Nadadores	CENTRO	6,335	711.74
05022	Nava	NORTE	27,928	904.86
05023	Ocampo	CENTRO	10,991	25,898.86
05024	Parras	SURESTE	45,401	10,517.78
05025	Piedras Negras	NORTE	152,806	473.24
05026	Progreso	CARBONIFERA	3,473	2,868.30
05027	Ramos Arizpe	SURESTE	75,461	6,692.50
05028	Sabinas	CARBONIFERA	60,847	1,966.33
05029	Sacramento	CENTRO	2,314	287.23
05030	Saltillo	SURESTE	725,123	5,562.31
05031	San Buenaventura	CENTRO	22,149	6,409.46
05032	San Juan de Sabinas	CARBONIFERA	41,649	798.67
05033	San Pedro	LAGUNA	102,650	7,080.25
05034	Sierra Mojada	CENTRO	6,375	7,873.54
05035	Torreón	LAGUNA	639,629	1,269.84
05036	Viesca	LAGUNA	21,319	4,357.68
05037	Villa Unión	NORTE	6,289	1,846.08
05038	Zaragoza	NORTE	12,702	7,919.85

Cuadro C.2**:** *Distribución de ingreso para Coahuila, información KL, Índice de Gini y promedio estandarizado*

CM	NOM	POB	F1	F2	F3	F4	F5	F6	F7	F8	IKL	IKL2	G	M
001	Abasolo	376	209	132	15	13	1	2	3	1	0.501	3.61	0.37	-1.05
002	Acuña	3107	3039	56	7	4	0	0	0	1	0.943	0.08	0.37	-0.59
003	Allende	1160	809	229	61	36	9	7	1	8	0.546	2.22	0.41	1.00
004	Arteaga	966	952	6	4	2	0	1	0	1	0.954	0.04	0.37	-0.40
005	Candela	508	503	3	1	0	0	0	0	1	0.969	0.03	0.41	-0.47
006	Castaños	1241	1060	143	22	6	7	1	1	1	0.746	0.90	0.35	-0.05
007	Cuatro Ciénegas	906	679	143	39	20	13	7	2	3	0.587	1.75	0.44	-0.44
008	Escobedo	815	653	140	15	4	1	1	0	1	0.709	1.35	0.34	-0.72

009	Francisco I. Madero	1258	1192	46	12	4	2	1	0	1	0.877	0.24	0.38	-0.86
010	Frontera	3500	2988	434	47	15	7	3	4	2	0.757	0.92	0.33	0.34
011	General Cepeda	845	801	31	7	2	1	0	2	1	0.877	0.24	0.41	-1.36
012	Guerrero	593	457	104	22	5	1	2	1	1	0.654	1.58	0.35	-0.43
013	Hidalgo	452	102	268	60	13	3	3	0	3	0.463	7.21	0.26	-0.08
014	Jiménez	735	632	70	19	8	3	1	0	2	0.738	0.85	0.35	-1.46
015	Juárez	442	282	135	12	7	3	1	1	1	0.573	2.80	0.30	-1.02
016	Lamadrid	464	412	32	12	4	2	0	0	2	0.773	0.65	0.37	-0.27
017	Matamoros	1285	1283	1	0	0	0	0	0	1	0.994	0.00	0.37	-0.17
018	Monclova	3278	3206	58	7	2	1	0	2	2	0.941	0.07	0.43	1.69
019	Morelos	928	666	179	56	16	3	5	2	1	0.585	2.04	0.37	0.71
020	Múzquiz	1112	1075	29	3	2	1	1	0	1	0.916	0.13	0.36	-0.09
021	Nadadores	1042	1029	8	3	1	0	0	0	1	0.962	0.03	0.37	-0.75
022	Nava	1196	1185	9	0	0	0	1	0	1	0.972	0.03	0.39	0.88
023	Ocampo	700	533	113	37	9	4	2	0	2	0.627	1.65	0.43	-0.06
024	Parras	1272	1032	187	29	13	5	3	1	2	0.694	1.24	0.39	-0.69
025	Piedras Negras	2632	2336	232	43	11	6	2	1	1	0.791	0.65	0.38	0.43
026	Progreso	923	918	4	0	0	0	0	0	1	0.983	0.01	0.34	-0.16
027	Ramos Arizpe	1720	1480	155	56	20	6	2	0	1	0.740	0.85	0.45	1.99
028	Sabinas	1193	1084	84	13	8	0	3	0	1	0.818	0.51	0.40	1.13
029	Sacramento	703	518	150	14	14	2	3	1	1	0.630	1.90	0.36	-0.35
030	Saltillo	6783	6633	125	16	4	2	1	1	1	0.942	0.08	0.42	1.77
031	San Buenaventura	1465	1223	192	34	10	2	1	2	1	0.728	1.06	0.35	-0.00
032	San Juan de Sabinas	1128	975	111	32	4	3	0	2	1	0.755	0.82	0.43	1.54
033	San Pedro	1329	1270	51	3	3	1	0	0	1	0.901	0.20	0.40	-1.36
034	Sierra Mojada	829	827	0	0	1	0	0	0	1	0.991	0.01	0.38	1.84
035	Torreón	7093	7076	9	1	2	1	1	1	2	0.990	0.00	0.45	1.50
036	Viesca	918	521	367	20	6	2	0	1	1	0.600	3.56	0.27	-1.72
037	Villa Unión	675	542	92	22	9	4	1	3	2	0.664	1.30	0.35	-0.63
038	Zaragoza	726	703	16	4	1	0	1	0	1	0.918	0.12	0.45	0.36

ÍNDICE ALFABÉTICO