Benemérita Universidada Autónoma de Puebla

Facultad de Ciencias Físico-Matemáticas

Agrupación de los estados de la República Mexicana basada en clústeres longitudinales a partir de información del sector turístico en el periodo de 1992 a 2019.

TESIS

que para obtener el grado de:

Maestra en Ciencias Matemáticas

presenta:

Arely Maldonado Azcona

Directores de Tesis :

Dr. Víctor Hugo Vázquez Guevara
Dr. Ramón Álvaraz Vaz

Puebla, Puebla. Diciembre 2023

A mi familia.

Agradecimientos

Agradezco a mis padres, antes que a nadie por la oportunidad de la vida y la del aprendizaje, el estudio y las enseñanzas que todo eso conlleva; me siento muy feliz por este nuevo logro en mi vida que también es suyo. A mis hermanas Fer y Tere por los momentos divertidos, los complicados y por los que vendrán en un futuro; no habría podidio lograrlo sin ustedes. A mi amiga Miri, tu amistad y apoyo incondicional han sido sido elementos clave en este recorrido. Desde luego al Dr. Víctor Hugo Vázquez Guevara, por su confianza, guía y paciencia durante el proceso, nada de esto habría sido posible sin su ayuda; además al Dr. Ramón Álvarez Vaz quien también aportó sus amplios conocimientos y calidad profesional y humana para la consecución de éste trabajo. También quiero dar las gracias a los miembros del jurado, los Doctores: Bulmaro Juárez Hernández, Hortensia Josefína Reyes Cervantes, Fernando Velasco Luna y Rei Israel Ortega Gutiérrez, por sus valiosas aportaciones. Finalmente, quiero agradecer al CONAHCYT por el apoyo económico durante toda la maestría.

Introducción

El crecimiento de la información disponible es exponencial, muchas aplicaciones y actividades del mundo actual generan una enorme cantidad de ella, por ejemplo las redes inteligentes, la videovigilancia, los sistemas financieros, etc. El estudio de datos a través del análisis de clústeres se ha desarrollado como un método eficaz para categorizar datos en grupos. El agrupamiento de dichos datos es usado principalmente como método de organización y/o resumen, con distintos fines, entre los que destacan la obtención de información útil para identificar características sobresalientes de los datos, o bien para identificar similitudes entre los elementos, lo que puede generar un aporte importante para la descripción de fenómenos de varias disciplinas.

La agrupación en clústeres es un tema muy versátil que intuitivamente consiste en organizar un conjunto de datos considerando patrones de similitudes y/o diferencias, para que estos puedan entenderse más fácilmente. En Encyclopedia Britannica [11] se define el análisis de clústeres como el conjunto de herramientas y algoritmos que se utilizan para clasificar diferentes objetos en grupos de tal manera que la similitud entre dos objetos sea máxima si pertenecen al mismo grupo y mínima en caso contrario. Este es un tema muy estudiado, por lo que se han propuesto muchas técnicas, tanto transversales como longitudinales [1]. La característica definitoria de un estudio de clústeres longitudinales es que cierta característica asociada con los individuos que son objeto de estudio se mide repetidamente a través del tiempo. Los estudios longitudinales contrastan con los estudios transversales en que los de tipo transversal miden un resultado único

para cada individuo [24]. El análisis de datos longitudinales es importante en los estudios correlacionales que buscan correspondencia entre observaciones de las mismas variables sobre algún periodo. Los estudios longitudinales permiten evaluar y estudiar los cambios en el tiempo de las variables de interés. Por ejemplo, se puede estudiar el turismo en la República Mexicana a través de técnicas de datos longitudinales para lograr una clasificación de los estados basados en información histórica del mismo, lo que permitiría construir bloques que podrían derivar en alianzas estratégicas entre ellos o bien diseñar estrategias promocionales comunes para detonar o potenciar el turismo, así como el diseño de políticas públicas comunes para estados que pertenezcan a un mismo clúster.

De acuerdo con la Organización Mundial del Turismo, el *turismo* es un fenómeno social, cultural y económico que supone el desplazamiento de personas a países o lugares fuera de su entorno habitual por motivos personales, profesionales o de negocios. El abordaje estadístico de este fenómeno en su aspecto económico puede hacerse desde el punto de vista de la demanda o desde el punto de vista de la oferta. La demanda se define en función del perfil de los visitantes (sexo, edad, nivel educativo, nivel de ingresos, etc.), medio de transporte utilizado, tipo de alojamiento utilizado, destinos elegidos, periodo de las estadías, finalidad o moivo de viaje, actividades desarrolladas en el lugar visitado y su impacto económico en las diferentes ramas de actividades. La oferta se define en función de las ramas de actividades relacionadas con la satisfacción del consumo turístico: hoteles y establecimientos de la estadía, inmuebles en alquiler para el turismo, restaurantes, cafés, transportes, agencias de viaje y operadores de turismo.

Según la última actualización del Sistema Nacional de Información Estadística del Sector Turismo de México-DataTur, con la llegada de 38 millones 326 mil turistas internacionales a México en 2022, nuestro país se posicionó en el sexto lugar de la clasificación mundial del la OMT (Organización Mundial del Turismo), por debajo de Turquía e Italia. Además se posicionó en la novena posición a nivel mundial por ingreso de divisas de visitantes internacionales, al sumar 28

mil 16 millones de dólares. Lo cual tuvo una significante participación en la actividad económica mexicana, pues el sector turístico en ese año aportó el 7.6 % del Producto Interno Bruto (PIB) a precios corrientes.

Debido a la significante aportación del turismo a la economía del país, se han hecho algunos estudios relacionados con el mismo. En [10], basados en un análisis de la relación que tienen el PIB real de México y el turismo internacional, a través de un modelo ARDL, se concluye que existe una relación positiva entre las tasas de crecimiento del PIB y la llegada de turistas internacionales, además se señala que este sector es muy sensible a los periodos de crisis. También, se afirma que, aunque la captación de turistass tiene un efecto positivo para la economía, se debe procurar que los vistantes sientan deseos de realizar desembolsos más cuantiosos, lo que podría lograrse poniendo en marcha una estrategia para aumentar el tiempo de permanencia en el país.

Por otro lado, en [21] realizaron un estudio a través de un análisis econométrico utilizando el modelo de cointegración de Johansen para examinar los factores que determinan la demanda del turismo en México. Los resultados obtenidos en este documento muestran que a largo plazo el precio relativo de los servicios turísticos es un factor determinante en la decisión de compra y que el mercado es sensible a las variaciones en el precios.

Otro trabajo notable en el estudio del sector turístico es la investigación hecha en [19]. En éste se propone un modelo de regresión lineal múltiple con las principales variables del turismo para conocer la tendencia de la variable del PIB de este sector. Se concluye que existe una correlación positiva entre el PIB del turismo con las variables *PIB de alojamiento*, *PIB de trasnporte aéreo*, *número de habitaciones* y *numero de turistas*. Por lo que se considera que es un acierto seguir apostando por el buen funcionamiento de la insdustria turística. También se menciona que la tendencia positiva que sigue el PIB del turismo inspira confianza para la llegada de una nueva IED (inversión extranjera que establece una participación prolongada en una empres o un control efectivo de su gestión) en

actividades correlacionadas con la actividad turística, principalmente en servicios de hotelería y restaurantes.

Desde otra perspectiva, en [8] se examinaron destinos turísticos de México desde una óptica cuantitativa que permitió evidenciar los sitios de concentración de esta actividad económica en el país. Se utilizaron cuatro tasas para evaluar el impacto de la intensidad del turismo extranjero en el terrritorio mexicano y se agruparon los municipios turísticos de México en ocho tipos mediante el método de tipificación probabilística.

El objetivo general del presente trabajo es clasificar a los estados de la República Mexicana en clústeres en función de la evolución del turismo en el periodo de enero de 1992 a diciembre de 2019. Para esto se discuten métodos de análisis de clústeres transversales y longitudinales. Lo cuál representa un trabajo novedoso, pues hasta ahora no se ha trabajado con un análisis estadísitico basado en clústeres longitudinales en el sector turístico en México.

Para lograr el objetivo se realizó un análisis de clústeres para datos longitudinales basados en las variabales "Total de visitantes" , "Total de habitaciones disponibles" y "Porcentaje de ocupación" , considerando a los 32 estados de la República Mexicana, las cuales están medidas mensualmente en el periodo de enero de 1992 a diciembre de 2019; estos datos fueron tomados de la página oficial de la **SECTUR** en la siguiente liga: *http://www.datatur.sectur.gob.mx/SitePages/InfTurxEdo.aspx#carousel-datatur*, mismos que se encuentran recopilados en 32 compendios (uno por cada estado) con 26 variables cada uno. Cabe mencionar, que en principio se había considerado trabajar con las variables "Total de visitantes", "Total de cuartos" y una tercera variable relacionada con la economía. En efectos de de cumplir con ese objetivo, se realizó la búsqueda de dicha información, pero no fue hallada, por lo que se procedió a solicitarla a INEGI y SECTUR; y se obtuvo como respuesta la no existencia de los datos solicitados. Además, es crucial decir que los datos seleccionados no estaban completos, de modo que se imputaron los valores faltantes con el promedio de los valores de la trayectoria a la que pertenece la

observación faltante.

Así, se trabajó con el estudio de la evolución individual y conjunta de las variables, para lo que se hizo uso de las librerías *kml* y *kml3d* [3] del software **R** [2], destinadas a trabajar con trayectorias simples y cojuntas respectivamente. Se obtuvieron agrupaciones de los 32 estados en el periodo de enero de 1992 a diciembre de 2019, así como en cuatro subperiodos, resultantes de seccionar el periodo ya mencionado, con respecto a los sexenios presidenciales que ocurrieron dentro de ese lapso.

Índice general

Capítulo 1

Clústeres Transversales

Según la RAE, "clúster" es la adaptación gráfica propuesta por la voz inglesa *cluster*, usada con cierta frecuencia en español como tecnicismo perteneciente a diversos ámbitos y que tiene el sentido general de "grupo de elementos similares o cercanos, o agrupados en función de alguna característica o variable". Pero, resulta que la definición no sólo puede ser difícil sino que incluso puede estar fuera de lugar, por lo que se ha sugerido que el último criterio para evaluar el significado es el juicio del valor del usuario. Si el uso de este término produce una respuesta del valor para el usuario, eso es todo lo que se necesita. Sin embargo, es posible que ninguna definición sea suficiente para todas las situaciones [1].

Estudiar acerca del análisis de clústeres es de mucha importancia pues tiene aplicaciones en muchas disciplinas.

Los clústeres transversales se caracterizan por trabajar sobre datos en los que se mide un resultado único para cada objeto o individuo. O sea sobre un conjunto de objetos que puede ser identificado biyectivamente con el conjunto $\Omega = \{1, 2, \cdots, n\}$ de n objetos descritos por p variables $X_1, X_2, \cdots, X_p$ medidas en un instante fijo t. La información de dichos datos puede ser resumida en una matriz de dimensión $n \times p$.

$$X := \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{12} & \cdots & x_{np} \end{pmatrix},$$

donde x_{ij} representa el valor de la variable j del objeto i.

1.1. Medidas de proximidad

Al crear clústeres, es de vital importancia conocer qué tan cerca están los objetos entre sí. La aplicación de análisis de clústeres, por lo general, no solo requiere del cálculo de distancias entre objetos, sino, también de establecer distancias entre grupos, y entre grupos y objetos.

Las medidas que cuantifican esta cercanía son conocidas como similitudes, distancias o disimilitudes. La intuición indica que dos objetos serán cercanos si la disimilitud o distancia entre ellos es pequeña o bien, si la similitud es grande.

Sea E un conjunto no vacío de n elementos $E = \{1, 2, \cdots, n\}$, $d : E \times E \to R^+ \cup \{0\}$ una función y $d_{ij} := d(i, j)$. Se dice que d es un índice de similitud, si para cada $i, j, k \in E$, d cumple con las propiedades 1 y 2.

1. $d_{ij} = 0$ si y sólo si $i = j$.
2. $d_{ij} = d_{ji}$.
3. $d_{ij} \leq d_{ik} + d_{jk}$.

Si además cumple con la propiededad 3, se dice que d es una distancia [7].

Es importante mencionar que algunas de las técnicas de análisis de clústeres

requieren de convertir la matriz X en la matriz de proximidades

$$D := \begin{pmatrix} d_{11} & & & \\ d_{21} & d_{22} & & \\ \vdots & \vdots & \vdots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{pmatrix}.$$

Existe una amplia gama de posibles medidas que permiten encontrar la similitud o distancia entre los individuos, las cuáles permiten la cosntrucción de la matriz de proximidades mencionada , algunas de estas serán presentadas a continuación. Se abordarán medidas sugeridas para datos que contienen variables categóricas y continuas.

1.2. Medidas para datos categóricos

Cuando se trabaja con datos donde todas las variables son categóricas, se utilizan especialmente las medidas de similitud [1]. Usualmente estas medidas están acotadas entre 0 y 1, aunque ocasionalmente se expresan como porcentajes en un rango de 0 % a 100 % [1]. Desde luego es muy fácil convertir una medida de similitud s en una medida de disimilitud d tomando por ejemplo , $d = 1 - s$.

A priori, se sabe que si dos objetos tienen un coeficiente de similitud igual a uno significa que ambos tienen valores idénticos para cada una de las variables. Mientras que un coeficiente de similitud igual o cercano a cero indica que los individuos difieren en un número grande de variables.

1.2.1. Medidas de similitud para datos binarios

El tipo más común de datos categóricos multivariantes es donde todas las variables son binarias [1]. A partir de una matriz de incidencia $X = (x_{ij})$ de orden $(n \times p)$ con n objetos y p variables binarias, la información acerca de la cercanía entre cualquier par de objetos, digamos $\boldsymbol{i}$=$(x_{i1}, \cdots, x_{ip})$ y $\boldsymbol{j}$=$(x_{j1}, \cdots, x_{jp})$,

puede representarse como una tabla de contingencia (Tabla 1.1), donde $a =$

Tabla 1.1: Tabla de contingencia.

		individuo i		
	resultados	1	0	Total
individuo j	1	a	b	a+b
	0	c	d	c+d
	Total	$a+c$	b+d	p=a+b+b+d

$|\{k|x_{ik}=1=x_{jk}\}|$ (es el número variables en las que los objetos i y j tienen el valor 1), $b = |\{k|x_{ik}=0\ y\ x_{jk}=1\}|$ (es el número de variables en las que el objeto i tiene valor 0 y el objeto j tiene valor 1), $c = |\{k|x_{ik}=1\ y\ x_{jk}=0\}|$(es el número de variables en las que el objeto i tiene valor 1 y el objeto j tiene valor 0) y $d = |\{k|x_{ik}=0=x_{jk}\}|$ (es el número variables en las que los objetos i y j tienen el valor 0). Note que pueden construirse $\frac{n(n-1)}{2}$ tablas de este tipo, las cuales ayudan a definir la distancia entre los objetos en función de a, b, c y d como [15]:

$$s_{ij} = \frac{a+\delta d}{\alpha a + \lambda(b+c) + \beta d}, \tag{1.1}$$

donde los parámetros δ, β, δ y λ son valores a elegir que determinarán distancias entre los objetos i y j, como se muestra en la siguiente (Tabla 1.2).

Tabla 1.2: Medidas de similitud para datos binarios.

Medida	δ	λ	α	β	fórmula
S1:Matching coefficient	1	1	1	1	$s_{ij} = \frac{a+d}{a+b+c+d}$
S2:Jaccard oefficient	0	1	1	0	$s_{ij} = \frac{a}{a+b+c}$
S3:Rogers and Tanimoto	1	2	1	1	$s_{ij} = \frac{a+d}{a+2(b+c)+d}$
S4:Sneath and Sokal	0	2	1	0	$s_{ij} = \frac{a}{a+2(b+c)}$
S5:Gower and Legendre	0	$\frac{1}{2}$	1	1	$s_{ij} = \frac{a+d}{a+\frac{1}{2}(b+c)+d}$
S6:Gower and Legendre	1	$\frac{1}{2}$	1	0	$s_{ij} = \frac{a}{a+\frac{1}{2}(b+c)}$

Observe también que entre mayores sean a y d, mayor será la similitud entre

los objetos i y j.

1.2.2. Medidas de similitud para datos categóricos con más de dos niveles

Los datos categóricos en los que las variables tienen más de dos niveles (por ejemplo, el tamaño: grande, mediano y pequeño) pueden estudiarse de manera similar a los datos binarios considerando a cada nivel de una variable como una sola variable binaria. Sin embargo, es posible que, en muchos casos este no sea una método atractivo [1]. Un método que es considerado más factible consiste en asignar el valor s_{ijk} que puede ser cero o uno asociado con la variable k ($k = 1, \cdots, p$), dependiendo de si los individuos i y j son iguales en esa variable, esto es,

$$s_{ijk} = \begin{cases} 0, & si \quad x_{ik} \neq x_{jk}; \\ 1, & si \quad x_{ik} = x_{jk} \end{cases} .$$

Luego, todos los valores se promedian sobre las p variables para obtener el coeficiente de similitud como sigue [15]:

$$s_{ij} = \frac{1}{p} \sum_{k=1}^{p} s_{ijk}. \tag{1.2}$$

1.2.3. Medidas de disimilitud y distancia para datos continuos

Si ahora la matriz de datos X, se construye a partir de los valores observados para n objetos y p variables continuas, los individuos a agrupar se pueden considerar como elementos en $\mathbb{R}^p$. En este caso, la proximidad entre los individuos generalmente se calcula mediante medidas de disimilitud.

Existe una gran variedad de medidas usadas para construir la matriz de distancias D en donde d_{ij} denota a la distancia entre los individuos i y j para este

tipo de datos. Las medidas más comunes se enlistan a continuación, aunque se pueden encontrar muchas más [1].

1. D1: Distancia Euclidiana

$$d_{ij} = \left[\sum_{k=1}^{p} w_k^2 (x_{ik} - x_{jk})^2 \right]^{1/2}. \tag{1.3}$$

   Donde x_{ik} y x_{jk} respresentan el valor de la observación de la $k-$ésima variable para los individuos i y j, respectivamente.

2. D2: Distancia de Manhattan

$$d_{ij} = \sum_{k=1}^{p} w_k |x_{ik} - x_{jk}|. \tag{1.4}$$

3. D3: Distancia de Minkowski

$$d_{ij} = \left(\sum_{k=1}^{p} w_k^r |x_{ik} - x_{jk}|^r \right)^{1/r} \qquad (r \geq 1). \tag{1.5}$$

4. D4: Distancia de Canberra

$$d_{ij} = \begin{cases} 0 & si\ x_{ik} = 0 = x_{jk}, \qquad para\ toda\ k. \\ \\ \sum\limits_{k=1}^{p} w_k |x_{ik} - x_{jk}| / \left(|x_{ik}| + |x_{jk}| \right) & para \qquad x_{ik} \neq 0\ o\ x_{jk} \neq 0. \end{cases} \tag{1.6}$$

5. D5: Correlación de Pearson

$$d_{ij} = \left(1 - \phi_{ij}\right)/2, \tag{1.7}$$

   con

$$\phi_{ij} = \sum_{k=1}^{p} w_k (x_{ik} - \overline{x}_i \cdot)(x_{jk} - \overline{x}_j \cdot) / \left[\sum_{k=1}^{p} w_k (x_{ik} - \overline{x}_i \cdot)^2 \sum_{k=1}^{p} w_k (x_{jk} - \overline{x}_j \cdot)^2 \right]^{1/2},$$

donde

$$\overline{x}_i \cdot = \sum_{k=1}^{p} w_k x_{ik} / \sum_{k=1}^{p} w_k.$$

6. D6: Separación angular

$$\delta_{ij} = (1 - \phi_{ij})/2, \quad con \quad \phi_{ij} = \sum_{k=1}^{p} w_k x_{ik} x_{jk} / \left(\sum_{k=1}^{p} w_k x_{ik}^2 \sum_{k=1}^{p} w_k x_{jk}^2 \right)^{1/2}. \tag{1.8}$$

Las medidas listadas están formuladas de modo que permitan una ponderación a través de los pesos $w_1, w_2, \cdots, w_n$. Aquí, se omite la cuestión de cómo elegir estos pesos pero puede consultarse en [1].

La medida más popular es la llamada distancia euclidiana que tiene la propiedad de que cada disimilitud d_{ij} puede interpretarse como una distancia física entre dos puntos en el espacio Euclidiano; la distancia de Manhattan también conocida como distancia del taxi, mide distancias en configuración rectilínea; la medida de disimilitud de Canberra es sensible para valores pequeños cercanos a $x_{ik} = 0 = x_{jk}$; comúnmente es considerada como una medida de disimilitud para datos binarios, para esto se puede dividir entre el número de variables, para asegurar un coeficiente de similitud dentro del intervalo $[0, 1]$ [1].

Las medidas $D5$ y $D6$ cuantifican la similitud entre dos individuos a través de la correlación entre las observaciones de dimensión p de cada objeto.

1.2.4. Medidas de proximidad entre clústeres

Existen; al menos, dos formas de definir proximidades entre clústeres, por un lado podría definirse mediante un resumen adecuado de las distancias entre los objetos de cada grupo. Por otra parte, cada clúster podría ser descrito por una observación representativa y la distancia entre grupos definida como la distancia entre las observaciones representativas.

Existe una gran variedad de medidas entre grupos que se calculan a partir de la matriz de distancias, D [1]. A saber, se pude considerar la menor disimilitud

entre individuos, uno en cada grupo, si por ejemplo, $I = \{i_1, i_2, \cdots, i_k\}$ y $J = \{j_1, j_2, \cdots, j_l\}$ representan a dos clústeres, con $k, l \in \mathbb{N}$, la distancia entre los grupos I y J será $D(I, J) = min\{d_{i_m, j_r}\}$. Esto se conoce como la distancia del vecino más cercano o enlace simple. Lo opuesto a la distancia del vecino más cercano es definir la distancia entre grupos como la distancia más grande entre dos individuos, uno en cada grupo, y esta distancia se conoce como la distancia del vecino más lejano o enlace completo. Otra opción es definir la distancia como el promedio de las distancias entre los individuos de ambos grupos.

Un método muy común para la construcción de medidas de similitud entre grupos para datos continuos, es sustituir a los valores de las variables en las fórmulas para medidas entre objetos, por los valores de los centroides (vector de medias) de los grupos. Por ejemplo, si un grupo A tiene como centroide al vector $\overline{\boldsymbol{X}}_{\boldsymbol{A}}$=$(\overline{x}_{A1}, \cdots, \overline{x}_{Ap})$ y B al vector $\overline{\boldsymbol{X}}_{\boldsymbol{B}}$=$(\overline{x}_{B1}, \cdots, \overline{x}_{Bp})$, la distancia euclidiana entre grupos puede ser definida como:

$$d_{AB} = \left[\sum_{k=1}^{p} (\overline{x}_{Ak} - \overline{x}_{Bk})^2 \right]^{1/2}. \tag{1.9}$$

Aunque para muchos autores, podrían ser más apropiadas las medidas que incorporan de alguna forma la variación dentro del grupo [1]. Una posibilidad es usar la distancia generalizada de Mahanlanobis´s, $\boldsymbol{D^2}$, dada por

$$\boldsymbol{D^2} = (\overline{\boldsymbol{X}}_{\boldsymbol{A}} - \overline{\boldsymbol{X}}_{\boldsymbol{B}})' \boldsymbol{W^{-1}} (\overline{\boldsymbol{X}}_{\boldsymbol{A}} - \overline{\boldsymbol{X}}_{\boldsymbol{B}}), \tag{1.10}$$

donde $\boldsymbol{W}$ es llamada matriz de covarianza agrupada para dos grupos.

1.3. Clústeres Jerárquicos

Entre los métodos más comunes de análisis de clústeres se encuentran los jerárquicos. Esta categoría se divide en dos subcategorías, los métodos aglomerativos, que actúan mediante una serie de fusiones de n individuos en grupos y los méto-

dos divisivos, que separan al grupo de n individuos en grupos más finos. En cada etapa de agrupamiento cada método opera con una matriz de proximidad.

Con este tipo de métodos, las divisiones o aglomeraciones son irreversibles, es decir, una vez que se han unido a dos individuos no pueden ser separados, y cuando se ha hecho una división, ésta no se puede deshacer.

Los métodos jerárquicos pueden representarse a través de un diagrama bidimensional conocido como dendrograma, donde los nodos representan los clústeres y las longitudes de los tallos representan las distancias a las que se unen o separan los grupos.

Debido a que los métodos jerárquicos terminan cuando los datos se reducen a un sólo grupo que contiene a todos los individuos o dividen al conjunto en n grupos de un elemento, se desea tener un número óptimo de clústeres. Hay una gran variedad de métodos para decidir dicho número. Uno de tales métodos consiste en hacer un corte horizontal a una determinada altura del dendograma, luego el número de ramas que sobrepasan en sentido ascendente sugiere el número de clústeres [1].

Además, existen otros métodos que ayudan a elegir el número de grupos en el análisis, llamadas **reglas de detención**, éstas se clasifican en dos categorías, **globales** y **locales**. Las reglas globales evalúan la bondad de la partición en g clústeres basándose usualmente en las variaciones dentro y entre grupos e identificando el valor g que optimiza la medida considerada. Las reglas de detención locales examinan si deben unirse dos grupos, basándose solamente en una parte de los grupos. A continuación se presentan algunas de tales reglas.

- **Calinski y Harabasz 1974:**

$$Pseudo\ F = \frac{\frac{tr(B)}{g-1}}{\frac{tr(W)}{n-g}} \tag{1.11}$$

donde W denota la matriz de dispersión dentro de los grupos, es decir,

$$W = \sum_{m=1}^{g} \sum_{l=1}^{n_m} (x_{ml} - \overline{x}_m)(x_{ml} - \overline{x}_m)^T, \tag{1.12}$$

con $\overline{x}_m$ el vector de dimensión p de las medias muestrales dentro del grupo m, g el número de grupos y n_m el número de objetos en el clúster m. Y B es la matriz de dispersión entre grupos, esto es,

$$B = \sum_{m=1}^{g} n_m (x_{ml} - \overline{x})(x_{ml} - \overline{x})^T. \tag{1.13}$$

Esta regla se comporta 'muy bien' en datos con estructura de grupos muy compactos y en presencia de distribuciones normales multivariadas. Empíricamente se han determinado algunas reglas que contribuyen a su utilización [17]:(1) Si el indicador crece monótonamente al crecer el número de grupos g, entonces no se puede determinar una estructura clara. (2) Si el indicador disminuye monótonamente al crecer el número de grupos g, entonces no se puede determinar claramente estructura de grupos, pero se puede decir que existe una estructura jerárquica. (3) Si el indicador crece, llega a un máximo y luego decrece, entonces la población presenta un número definido de grupos en ese máximo [17].

- **Pseudo t^2:** Es un indicador útil de grupos, pero no se distribuye como una variable aleatoria t de student. Es una regla de detención local que analiza en cada paso la unión de dos grupos L y G, donde $tr(W_L)$ denota la traza de la matriz de dispersión dentro del clúster L; y n_L, n_G denotan la cantidad de individuos en los clústeres L y G respectivamente.

$$Pseudo\ t^2 = \frac{tr(W_{GL}) - tr(W_G) + tr(W_L)}{\frac{tr(W_G)+tr(W_L)}{n_G+n_L-2}}. \tag{1.14}$$

Se trata de determinar en cada paso si la unión de dos grupos implica un

incremento en la suma de cuadrados residuales (variación intragrupos, o variación entre los grupos) que la misma deba ser desechada.

Si el análisis se realiza desde los n individuos a un único grupo final (de arriba hacia abajo) entonces: si en el paso $k+1$ el indicador (1.14) tiene un valor muy pequeño con respecto al nivel del paso k entonces es conveniente quedarse con $k+1$ grupos en lugar de unir los grupos que se podían unir en esa instancia. Esto es, el incremento de la heterogeneidad de unir esos grupos es muy grande y por tanto no es conveniente hacerlo.

Si el análisis se realiza desde un único grupo hacia los n individuos (de abajo hacia arriba) entonces: Si al pasar de k grupos a $k+1$ el indicador (1.14) tiene una caída 'importante' (en k el indicador es más grande que en $k+1$) es conveniente quedarse con k + 1 grupos. Esto es, el incremento en la heterogeneidad de unir esos grupos es muy grande y por tanto no es conveniente unirlos.

- **Gráfico de Silueta** Los gráficos de Silueta pueden ser usados para: (1) Seleccionar el número de clústers. (2) Evaluar cuan bien han sido asignadas las observaciones en los grupos. Sea s_i el ancho de la silueta de la observación i

$$s_i = \frac{b_i - a_i}{max(b_i, a_i)}, \tag{1.15}$$

donde a_i es la distancia promedio de la observación i a los objetos de su grupo y b_i la distancia promedio de la observación i a los objetos de los demás grupos.

Las observaciones con ancho de silueta grande están bien agrupadas mientras que aquellas con ancho de silueta moderada tienden a estar ubicadas en medio de distintos grupos [17]. Con el ancho de silueta promedio definido como el promedio de todas las observaciones, para un número de grupos g:

$$\overline{s} = \frac{1}{n} \sum_i s_i \tag{1.16}$$

se sugiere considerar el número de grupos g para el cual el ancho de silueta promedio sea la mayor posible [16].

1.3.1. Métodos aglomerativos

Los métodos jerárquicos aglomerativos generan una serie de particiones, la primera consta de n clústeres individuales y la última consiste de un solo grupo que contiene a todos los individuos.

Como ya se vio anteriormente (Sección 1.1), existen varias medidas de proximidad entre grupos, cada una de las cuales deriva técnicas de clústeres jerárquicos diferentes. Las técnicas más destacadas dentro de este método son la de enlace simple, el enlace completo, el enlace centroide, el enlace promedio y el método de Ward [1]. Como se ha mencionado, en el método de enlace simple la medida se define como la distancia más pequeña entre los clústeres. Mientras que en el enlace completo se define como la distancia máxima. El enlace centroide considera la distancia calculada entre los centroides de dos grupos. Finalmente,en el método de Ward, la fusión de dos clústeres está basada en el valor de la suma de errores al cuadrado, el objetivo en cada paso es minimizar el incremento de la suma de errores al cuadrado total, E, dentro de los clústeres, dada por

$$E = \sum_{m=1}^{g} E_m, \tag{1.17}$$

con

$$E_m = \sum_{l=1}^{n_m} \sum_{k=1}^{p_k} \left(x_{ml,k} - \overline{x}_{m,k}\right)^2, \tag{1.18}$$

donde $\overline{x}_{m,k} = \frac{1}{n_m} \sum_{l=1}^{n_m} x_{ml,k}$ y $x_{ml,k}$ es el valor de la k-ésima variable $k = 1, \cdots, p$ en el objetos l-ésimo $l = 1, \cdots, n_m$ en el clúster m; $m = 1, \cdots, g$.

Everitt, Landau, et al. [1] mencionan que las medidas de distancia entre grupos utilizadas en algunas técnicas de agrupación jerárquica pueden estar contenidas, bajo la elección adecuada de los parámetros, dentro de la fórmula de recurrencia

de Lance y Williams que calcula la distancia entre un grupo k y un grupo (ij) formado por la unión de los grupos i y j, y d_{ij} es la distancia entre dichos grupos:

$$d_{k(ij)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma \mid d_{ki} - d_{kj} \mid . \tag{1.19}$$

El enlace simple, por ejemplo, corresponde a los valores $\alpha_i = \frac{1}{2} = \alpha_j$; $\beta = 0$ y $\gamma = -\frac{1}{2}$.

$$d_{k(ij)} = \frac{1}{2} d_{ki} + \frac{1}{2} d_{kj} - \frac{1}{2} \mid d_{ki} - d_{kj} \mid . \tag{1.20}$$

Como resumen de lo anterior, se tiene el diagrama en la figura 1.1 ilustra la forma de ejecutar el método jerárquico aglomerativo.

A continuación se ilustran dos técnicas jerárquicas con ejemplos particulares, enlace simple y enlace centroide. La primera require unicamente de una matriz de distancias, por ejemplo, si se considera la matriz D_1, se procede de la siguiente manera:

$$D_1 = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{bmatrix} 0.0 & & & & \\ 1.2 & 0.0 & & & \\ 3.6 & 5.3 & 0.0 & & \\ 1.8 & 4.6 & 2.0 & 0.0 & \\ 6.0 & 2.5 & 4.0 & 2.5 & 0.0 \end{bmatrix}.$$

La entrada distinta de cero más pequeña en la matriz D_1 es la distancia entre los individuos 1 y 2, por lo que se unen para formar un grupo con dos individuos. Las distancias entre este nuevo grupo y los otros tres individuos se calcula como

$$\begin{aligned} d_{(12)3} &= min(d_{13}, d_{23}) = d_{13} = 3.6, \\ d_{(12)4} &= min(d_{14}, d_{24}) = d_{14} = 1.8, \\ d_{(12)5} &= min(d_{15}, d_{25}) = d_{25} = 2.5. \end{aligned}$$

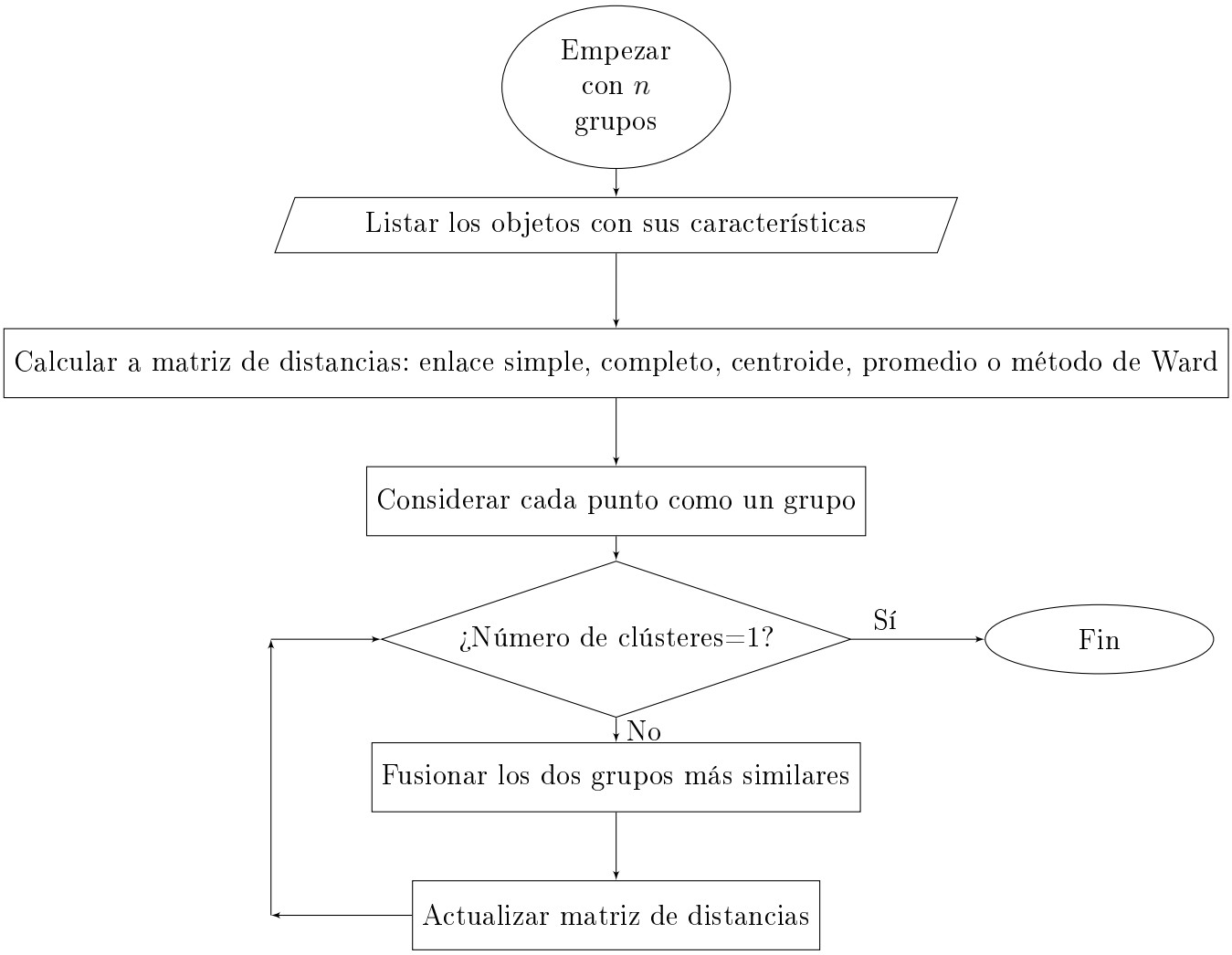


Figura 1.1: Diagrama de flujo de los métodos jerárquicos aglomerativos.

Ahora se puede construir una nueva matriz cuyas entradas representan valores que miden distancias entre individuos, y entre individuos y clústeres:

$$D_2 = \begin{array}{c} (12) \\ 3 \\ 4 \\ 5 \end{array} \begin{bmatrix} 0.0 & & & \\ 3.6 & 0.0 & & \\ 1.8 & 2.0 & 0.0 & \\ 2.5 & 4.0 & 2.5 & 0.0 \end{bmatrix}.$$

La entrada más pequeña en D_2 es la distancia entre el grupo (12) y el individuo 4, por lo que ahora forman un segundo grupo y se deben hallar nuevas distancias:

$$\begin{aligned} d_{(124)3} = &\quad min(d_{13}, d_{23}, d_{43}) &&= d_{13} &&= 4.0, \\ d_{(124)5)} = &\quad min(d_{15}, d_{25}, d_{45}) &&= d_{25} = d_{45} &&= 2.5, \\ d_{(35)} = &\quad 4.0. \end{aligned}$$

De lo que resulta una nueva matriz de distancias:

$$D_2 = \begin{array}{c} (124) \\ 3 \\ 5 \end{array} \begin{bmatrix} 0.0 & & \\ 2.0 & 0.0 & \\ 2.5 & 4.0 & 0.0 \end{bmatrix}.$$

Aquí, la entrada más pequeña es $d_{(124)3}$, entonces el individuo 3 se agrega al clúster que contiene a los individuos 1, 2 y 4. Finalmente se combinan los grupos que contienen a los individuos 1, 2, 4, 3 y al individuo 5.

Para ilustrar el método del enlace centroide, se trabajará con el conjunto de datos bivariados presentado en la Tabla 1.3 :

Al elegir la distancia euclidiana como la medida para trabajar en este caso, se obtiene la siguiente matriz de distancias:

Tabla 1.3: Datos para ilustrar el ejemplo usando el enlace centroide.

Objeto	Variable 1	Variable 2
1	1.0	1.0
2	2.0	2.0
3	2.0	3.0
4	5.0	3.0
5	7.0	6.0

$$C_1 = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{bmatrix} 0.00 & & & & \\ 1.41 & 0.00 & & & \\ 2.23 & 1.10 & 0.00 & & \\ 4.47 & 3.16 & 3.00 & 0.00 & \\ 7.81 & 6.40 & 5.83 & 3.60 & 0.00 \end{bmatrix}.$$

Examinando dicha matriz se observa que la distancia más pequeña es la de los objetos 2 y 3, por lo que estos objetos se unen para formar un nuevo grupo. A continuación se calcula el centroide para el nuevo grupo y después la nueva matriz de distancias:

$$C_2 = \begin{array}{c} 1 \\ (23) \\ 4 \\ 5 \end{array} \begin{bmatrix} 0.00 & & & \\ 1.80 & 0.00 & & \\ 4.47 & 3.04 & 0.0 & \\ 7.81 & 6.10 & 3.60 & 0.00 \end{bmatrix}.$$

Ahora la entrada más pequeña en C_2 es la distancia entre los clústeres (23) y (1), por lo que se unen para formar el grupo (123). Como antes, se calcula el centroide

para el nuevo grupo y después la nueva matriz de distancias:

$$C_3 = \begin{array}{c} (123) \\ 4 \\ 5 \end{array} \begin{bmatrix} 0.0 & & \\ 3.71 & 0.0 & \\ 6.95 & 3.60 & 0.0 \end{bmatrix}.$$

En C_3 la entrada más pequeña es la distancia entre los clústeres (123) y (4) por lo que se unen en un solo clúster de 4 miembros. La etapa final consiste en unir los dos últimos grupos en uno solo.

El dendrograma que ilustra el proceso y las particiones producidas en cada etapa se muestran en la figura 1.2.

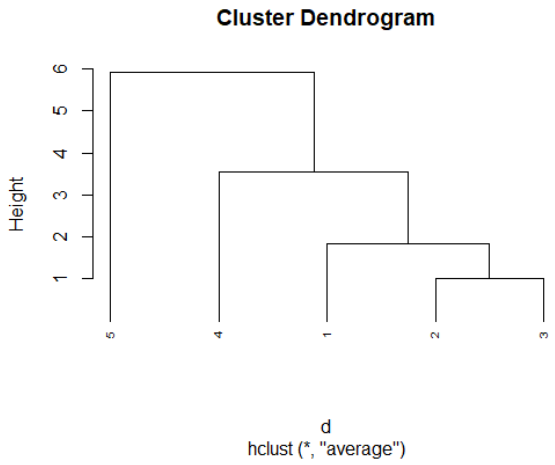


Figura 1.2: Dendograma que resulta de trabajar los datos de la Tabla 1.3 y haciendo uso de enlace centroide.

1.3.2. Métodos Divisivos

Los métodos divisivos actúan de forma opuesta a los métodos aglomerativos ya que, comienzan con un grupo de n elementos y se divide secuencialmente hasta obtener n grupos con un solo elemento. Sin embargo, comúnmente se prefieren los métodos aglomerativos debido a su eficiencia computacional, pues si se consideran los $2^n - 1$ divisiones posibles en dos subgrupos de un grupo de n objetos, los métodos divisivos tienden a volverse complicados. En consecuencia, se han sugerido métodos que no consideren todas las biparticiones posibles en cada paso. Cuando se trabaja con datos que consisten de p variables binarias se cuenta con los métodos divisivos monotéticos, el término *monotético* se refiere al uso de una sola variable en la cual basar la división en una etapa determinada, lo que quiere decir que estos métodos dividen a los grupos de acuerdo a la presencia o ausencia de alguna de las p variables. Por otro lado los métodos *politéticos* utilizan todas las variables simultáneamente en cada etapa.

La elección de la variable al trabajar con los métodos monotéticos puede depender de un criterio de inercia o de la asociación con otras variables [1]. A continuación se presenta una técnica de análisis de clústeres relacionada con el primer caso.

DIVCLUS-T

DIVCLUS-T es un método biparticional donde la elección de la bipartición (relacionada con una pregunta definida sobre una variable) y del clúster a dividir en cada paso se basa en la optimización de un criterio de inercia dentro de los clústeres [18]. Dada la partición $P_k = (C_1, C_2, \cdots, C_k)$, este método divide el clúster C_l que ayuda a encontrar una partición P_{k+1} que contenga $k+1$ clústeres y minimice la inercia dentro de estos. Considerando la matriz X y el vector $w = (w_1, \cdots, w_n)$, la medida de homogeneidad del clúster C_l es la inercia definida como $I(C_l) = \sum_{i \in C_l} w_i d^2 \left(x_i, g(C_l)\right)$ en donde d es la distancia Euclidiana y $g(C_l)$ el

centroide de C_l definido por $g(C_l) = \sum_{i \in C_l} \frac{w_i}{\sum_{k \in C_l} w_k} x_i$. Entonces una medida de adecuación para la partición P_k es la inercia total dentro de los clústeres, determinada como la suma de la inercia de todos los clústeres: $W(P_k) = \sum_{l=1}^{k} I(C_l)$.

En este método, para evitar considerar todas las biparticiones de un clúster, se utilizan preguntas binarias en cada paso, las cuales se definen sobre las p variables de la matriz de datos X. Una pregunta binaria en la variable numérica X_j es definida como "¿$X_j \leq c$?", con $c \in \mathbb{R}$, la cual divide al clúster C_l ($l = 1, \cdots, k.$) en los clústeres $A_l = \{i \in C_l \mid x_{ij} \leq c\}$ y $\overline{A}_l \{i \in C_l \mid x_{ij} > c\}$. Observe que, en este caso, si n_l denota el tamaño del clúster C_l, la cantidad de posibles preguntas binarias es infinita pero inducen solo $n_l - 1$ posibles biparticiones porque para los valores ordenados $x_{1j}, \cdots x_{ij}, \cdots, x_{n_l j}$, la pregunta ¿$X_j \leq c$? induce la misma bipartición $(\{1, \cdots, i\}, \{i+1, n_l\})$ para valores de c entre las observaciones consecutivas x_{ij} y $x_{i+1\ j}$, por lo que se asocia una única pregunta a cada bipartición $(\{1, \cdots, i\}, \{i+1, \cdots, n_l\})$ tomando a c como el punto medio entre dos observaciones consecutivas en cada partición. Pero si las variables son categóricas, una pregunta binaria es definida como "¿$X_j \in M$?", donde M denota el subconjunto de categorías de X_j, en este caso C_l será dividido en los subclústeres $A_l = \{i \in C_l \mid x_{ij} \in M\}$ y $\overline{A}_l = \{i \in C_l \mid x_{ij} \notin M\}$. En este caso, si q_j denota el número de categorías de una variable, la cantidad de posibles preguntas binarias , y por lo tanto el número de biparticiones es a lo más $2^{q_j - 1} - 1$.

DIVCLUST-T seleccionará entre todas las biparticiones $(A_l, \overline{A}_l)$ de C_l inducidas por todas las preguntas binarias sobre todas las variables, la que proporcione como resultado la menor inercia $W((A_l, \overline{A}_l))$ dentro de los clústeres de la bipartición $P_{C_l} = \left\{A_l, \overline{A}_l\right\}$, o, equivalentemente, la que maximice la inercia entre los clústeres de dicha bipartición, $B(A_l, \overline{A}_l) = \frac{\mu(A_l)\mu(\overline{A}_l)}{\mu(A_l)+\mu(\overline{A}_l)} d^2(g(A_l), g(\overline{A}_l))$, donde $\mu(C_l) = \sum_{i \in C_l} w_i$ [18].

Finalmente, el algoritmo elegirá dividir el clúster C_l de la partición P_k (de acuerdo con el algoritmo biparticional anterior) cuya división produce una nueva partición P_{k+1} con inercia intra clústeres mínima.

Es claro que, con este método, el conjunto de clústeres obtenidos después de $k-1$ biparticiones sucesivas es una jerarquía binaria, cuya definición puede encontrarse en [18]. Dado que no siempre se requiere continuar con las divisiones hasta que todos los grupos contengan un solo elemento, en este caso, la jerarquía que resulta se considera como una jerarquía parcial de modo que los singletones son los k clústeres de la partición P_k obtenida en el último paso del método. En esta jerárquía la altura de un grupo C_l dividido en dos subgrupos A_l y $\overline{A}_l$ está indexada por h tal que:

$$h(C_l) = B(A_l, \overline{A}_l) = \frac{\mu(A_l)\mu(\overline{A}_l)}{\mu(A_l) + \mu(\overline{A}_l)} d^2(g(A_l), g(\overline{A}_l)). \tag{1.21}$$

Así, DIVCLUS-T elige dividir el grupo C_l que maximice el valor de $h(C_l)$. Luego, como $W(P_{k+1}) = W(P_k) - h(C_l)$, maximizar $h(C_l)$ garantiza la mínima inercia dentro de los clústeres de la partición $P_{k+1} = P_k \cup \{A_l, \overline{A}_l\} - \{C_l\}$ [18].

Cabe mencionar que la ventaja que destaca sobre este método es la interpretación directa de los clústeres pues la jerarquía se puede leer como un árbol de decisión.

La Figura 1.3 muestra la descripción general del algoritmo:

Más precisamente, para llevar a cabo el proceso, DIVCLUS-T realiza los siguientes pasos secuencialmente:

1. Selecciona la mejor bipartición $(A_l, \overline{A}_l)$. Es decir, selecciona la bipartición de máxima inercia entre clústeres (definida anteriormente) entre las biparticiones inducidas por todas las preguntas binarias sobre todas las p variables $X_1, \cdots, X_p$.

2. Elige en la partición P_k el grupo C_l que se dividirá, de tal manera que la nueva partición P_{k+1} minimice la inercia dentro de los grupos.

3. El proceso se repite, si se desea, hasta obtener grupos con un solo individuo.

Para el segundo caso, como ya se mencionó, en lugar de la homogeneidad del conglomerado la variable en cada paso se elige de acuerdo con su asociación general

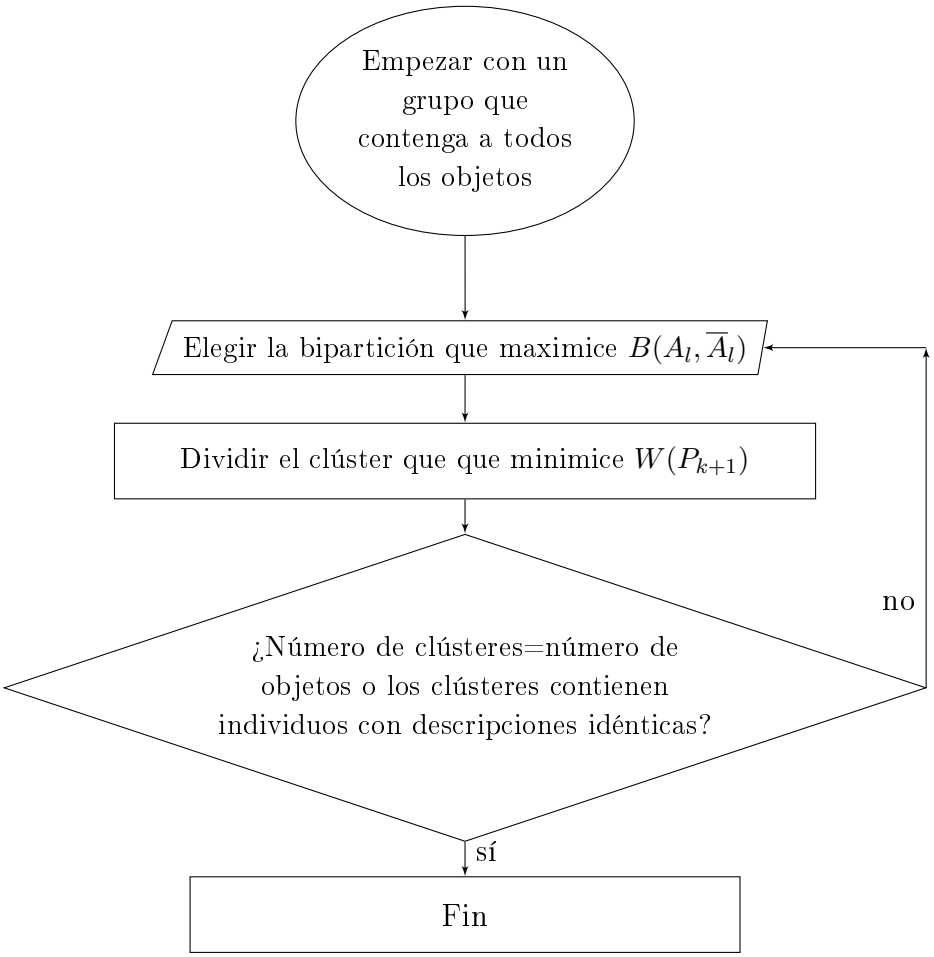


Figura 1.3: Diagrama de flujo del método DIVCLUS-T.

con todas las variables restantes (análisis de asociación), se puede proceder de la siguiente manera: para un par de variables X_i y X_j las frecuencias observadas pueden resumirse en una tabla como la siguiente:

La idea es que se identifique la variable que tiene mayor asociación con el resto de las variables. Los objetos se dividen en dos grupos de acuerdo a la presencia o ausencia de dicha variable cuya asociación con los demás es un máximo.

Las medidas más comunes de asociación son las siguientes:

$$|ad - bc|, \tag{1.22}$$

$$(ad - bc)^2 n/[(a + b)(a + c)(b + d)(c + d)], \tag{1.23}$$

Tabla 1.4: Pseudocódigo del método DIVCLUS-T para datos continuos.

Pseudocódigo del método DIVCLUS-T para datos continuos.

Entrada: Inicializar $k = 1$
Inicializar la partición $P_k = \{\{1, 2, \cdots, n\}\}$
Mientras (las divisiones no den grupos con un solo individuo o grupos de individuos con descripciones idénticas) **hacer**
Para $j = 1$ a p
para $l = 1$ a k
para $c \in D_l$
calcular $B(A_l, \overline{A}_l)$
fin para
fin para
fin para
$r =$Número de posibles biparticiones
mientras $W(P_{k+1})$ no sea mínimo **hacer**
para $j = 1$ a p
para $r = 1$ a k
revisar el valor de $W(P_{k+1}) = B(A_l, \overline{A}_l)$
fin para
fin para
fin mientras
Actualizar $P_{k+1} = P_k \cup \{A_l, \overline{A}_l\} - \{C_l\}$
$k = k + 1$ **Fin mientras**

Tabla 1.5: Tabla de frecuencias.

	X_i	
X_j	1	0
1	a	b
0	c	d

$$\sqrt{(ad-bc)^2 n/[(a+b)(a+c)(b+d)(c+d)]}, \tag{1.24}$$

$$(ad-bc)^2/[(a+b)(a+c)(b+d)(c+d)]. \tag{1.25}$$

En contraste, los métodos divisivos politéticos son más parecidos a los métodos aglomerativos, en el sentido que se usan todas las variables simultáneamente y además se trabaja con la matriz de proximidad.

En cada paso el algoritmo desea dividir a un conjunto en dos, pero considerar todas las divisiones posibles representa un gran trabajo en este tipo de métodos, por lo que se hace a través de un proceso iterativo. Así que primero se procede encontrando el objeto que está más alejado de los demás dentro de un grupo, el grupo principal, y usándolo como semilla para formar el grupo disidente. Después, cada uno de los objetos del grupo principal es considerado para entrar en el grupo separado, esto es, se busca el objeto que es más diferente a todos los demás del clúster principal y que está más cerca del clúster disidente, y se mueve a este último [1]. El siguiente diagrama (Figura 1.4) y ejemplo ilustran la forma de ejecutar dicho método:

Considere la siguiente matriz de distancias para siete individuos:

$$D_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{bmatrix} 0 & & & & & & \\ 13 & 0 & & & & & \\ 27 & 18 & 0 & & & & \\ 15 & 25 & 10 & 0 & & & \\ 16 & 26 & 21 & 9 & 0 & & \\ 7 & 28 & 32 & 17 & 38 & 0 & \\ 19 & 15 & 12 & 11 & 34 & 6 & 0 \end{bmatrix}$$

Considerando la distancia promedio de cada uno de los elementos a los restantes, cuyos valores se reflejan en la matriz de distancias D_1, se concluye que el elemento cuya distancia es mayor es el individuo 5, por lo que es el elegido para iniciar el grupo disidente, resultando como clústeres iniciales (5) y $(1, 2, 3, 4, 6, 7)$. Ahora se calcula la distancia promedio de cada individuo del clúster principal al resto y la distancia de cada individuo de dicho grupo al clúster nuevo. Intuitivamente esto indicará qué tan lejos está de cada uno de los individuos del grupo principal y del grupo disidente, respectivamente. Luego, se encuentra la diferen-

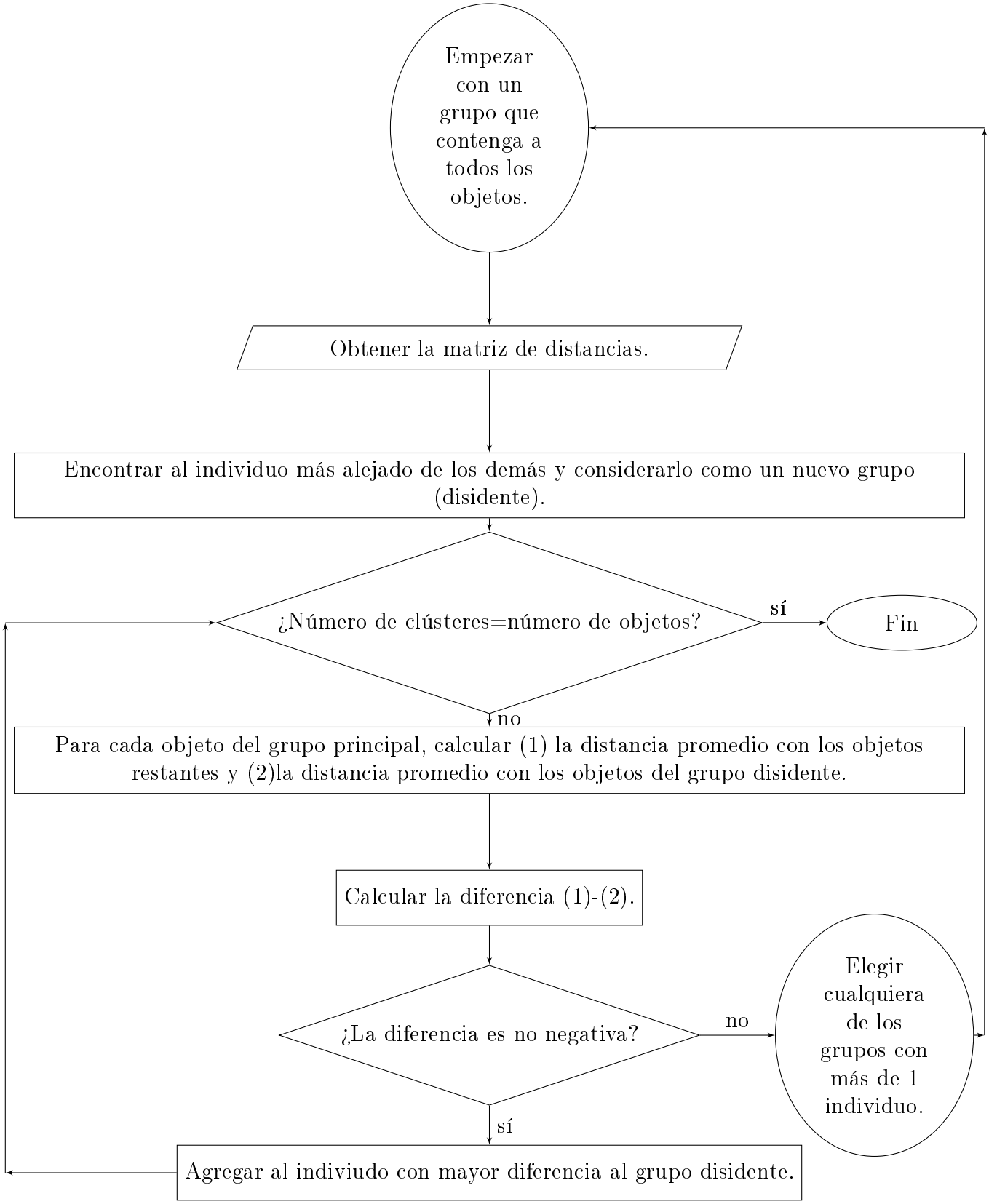


Figura 1.4: Diagrama de los métodos divisivos politéticos.

cia entre estas distancias, pues con esto se determinará cuál individuo se moverá al grupo disidente. Note que una diferencia negativa señala que un individuo es más disímil del grupo separado que del grupo principal, por lo que cuando esto ocurra se descartará la separación de este objeto. Los resultados obtenidos para este ejemplo se encuentran en la siguiente tabla (Tabla 1.6) :

Tabla 1.6: Tabla de resultados en el paso 2.

Individuo en el grupo principal	*Distancia promedio al grupo disidente (A)*	*Distancia promedio en el grupo principal (B)*	B-A
1	16	16.2	0.2
2	26	19.8	-6.2
3	21	19.8	-1.2
4	9	15.6	6.6
6	38	18.0	-20
7	34	12.6	-21.4

En este caso la diferencia más grande es la del individuo 4, por lo tanto se une al grupo que contiene el elemento (5), resultando los grupos (45) y (12367). Luego, repitiendo el proceso anterior se obtienen los valores de la Tabla 1.7.

Tabla 1.7: Tabla de resultados el paso 3.

Individuo en el grupo principal	*Distancia promedio al grupo disidente (A)*	*Distancia promedio en el grupo principal (B)*	B-A
1	15.50	16.50	1
2	25.50	18.50	-7
3	15.50	22.25	6.75
6	27.50	18.25	-9.25
7	22.50	13.00	-9.5

Así que ahora el individuo 3 se une al grupo disidente obteniendo los nuevos grupos (346) y (1267). Y repitiendo una vez más el proceso anterior se genera la Tabla 1.8.

Como ahora todas las diferencias son negativas, ningún elemento se puede añadir al clúster anterior. Por lo que, si se desea, el proceso se continúa en cada subgrupo por separado. Se puede comenzar con la división del grupo principal

Tabla 1.8: Tabla de resultados el paso 4.

Individuo en el grupo principal	*Distancia promedio al grupo disidente (A)*	*Distancia promedio en el grupo principal (B)*	B-A
1	19.30	13.00	-6.30
2	23.00	18.66	-4.34
6	29.00	13.66	-15.34
7	19.00	13.33	-5.67

(1267). Por los cálculos anteriores se comienza por separar al individuo 2. A continuación se calcula la distancia entre 2 y cada uno de los individuos del grupo (167), y la distancia de cada individuo de (167) a los elementos del mismo. Como antes, repitiendo el proceso se obtienen los siguientes resultados (Tabla 1.9):

Tabla 1.9: Tabla de resultados el paso 5.

Individuo en el grupo principal	*Distancia promedio al grupo disidente (A)*	*Distancia promedio en el grupo principal (B)*	B-A
1	13.00	13.00	0
6	28.00	6.50	-21.5
7	15.00	12.50	-2.5

Observe que para el individuo 2 la diferencia es cero, es decir, este objeto es tan disímil del grupo principal como del grupo disidente, pero como las diferencias restantes son negativas, entonces el elemento elegido para separar del clúster principal es el 1, con lo que el individuo 1 se suma al individuo 2.

Se investigará ahora si algún otro individuos se une al clúster (12):

Tabla 1.10: Tabla de resultados el paso 6.

Individuo en el grupo principal	*Distancia promedio al grupo disidente (A)*	*Distancia promedio en el grupo principal (B)*	B-A
6	17.50	6.00	-11.15
7	17.00	6.00	-11

Por los resultados descritos en la TAbla 1.10, se concluye que no se agrega ningún otro individuo. Es decir, el clúster (1267) queda separado como (12) y

(67).

El proceso se puede continuar descomponiendo los clústeres restantes.

1.4. Clústeres basados en Criterios de optimización

Los métodos en este tipo de clústeres producen una partición de los objetos minimizando o maximizando algún criterio, es decir, a cada partición de los n individuos se asocia un índice numérico que mide la bondad de la partición. A diferencia de los métodos jerárquicos, en este tipo de método sí se admite la reasignación de los individuos. Por lo general en algunos de estos se supone que el número de grupos ya ha sido fijado, es decir, el investigador necesita estimar el número óptimo de clústeres en los que deben ser divididos los datos. En este capítulo, primero se describe un par de criterios que ayudan a elegir el mejor número de clústers y después se estudian algunos los índices en los que estarán basadas las particiones. En la sección 1.3 ya se han descrito criterios usados para la elección del número de clústeres usados principalmente en los métodos jerárquicos, pero aquí se abordarán más técnicas que se aplican igualmente bien a los métodos de optimización que los métodos de agrupación jerárquica [1].

Un criterio para definir cuándo es o no conveniente dividir a un clúster, digamos al clúster C_m, en dos subclústeres es descrito a continuación. Se comparan la suma de las distancias al cuadrado entre los individuos y el centroide del grupo, la cual es denotada por $J_1^2(C_m)$, con la suma de las distancias al cuadrado dentro del grupo cuando el grupo se divide óptimamente en dos, representada como $J_2^2(C_m)$. La hipótesis nula de que el clúster es homogéneo debe rechazarse y por lo tanto el clúster debe dividirse, si

$$L(C_m) = \left(1 - \frac{J_2^2}{J_1^2} - \frac{2}{\pi p}\right)\left\{\frac{n_m p}{2[1 - 8/(\pi^2 p)]}\right\}^{\frac{1}{2}} \tag{1.26}$$

excede el valor crítico de una distribución normal estándar. Como antes, p es el número de variables y n_m el número de individuos en el clúster m.

Otro criterio que ayuda a elegir cuándo es oportuno dividir un clúster en dos, que también involucra la suma de las distancias al cuadrado es la siguiente: Si ahora S_g^2 denota las desviaciones al cuadrado de los centroides en la muestra. Entonces se afirmará que una división de los n objetos en g_2 clústeres es significativamente mejor que una división en g_1 grupos ($g_2 > g_1$) si la estadística de prueba

$$F(g_1, g_2) = \frac{\left(S_{g_1}^2 - S_{g_2}^2\right) / S_{g_2}^2}{[(n-g_1)/(n-g_2)](g_2/g_1)^{2/p} - 1} \tag{1.27}$$

supera el valor crítico de una distribución F con $p(n-g_1)$ y $p(n-g_2)$ grados de libertad.

Además de estos métodos para la elección del número de clústeres, existen muchos más que no serán abordados aquí.

Por otro lado, para elegir la mejor manera de agrupar a los elementos, se presentan criterios que hacen uso de la matriz de disimilitudes, criterios exclusivos para variables continuas y además algoritmos para optimizar dichos criterios.

Los criterios basados en la matriz de disimilitudes hacen uso de los conceptos de homogeneidad y separación, pues se desea obtener una partición que genere grupos con estructura cohesiva, esto es, que minimice la falta de homogeneidad dentro de los clústeres y además se logre heterogeneidad entre ellos. Para esto, consideremos las medidas de adecuación de la Tabla 1.11:

Donde $\delta_{ql,k\nu}$ son las entradas de la matriz de disimilitud y denotan la disimilitud entre el individuo l en el grupo q y el individuo ν en el grupo k.

Observe que $h_1(m), h_2(m), h_3(m)$ miden la heterogeneidad del grupo C_m, mientras que $i_1(m)$ y $i_2(m)$ cuantifican la separación. Una vez elegido el índice que mide la homogeneidad o la separación de un grupo, se desea tener criterios que cuantifiquen la bondad de las agrupaciones, mediante una agregación adecuada, por

Tabla 1.11: Medidas de adecuación.

Medida	*índice*
Falta de homogeneidad	$h_1(C_m) = \sum_{l=1}^{n_m} \sum_{\nu=1,\nu\neq l}^{n_m} (\delta_{ml,m\nu})^r$
Falta de homogeneidad	$h_2(C_m) = \max_{\substack{l=1\ldots,n_m \\ \nu=1\ldots,n_m \\ \nu\neq l}} [(\delta_{ml,m\nu})^r]$
Falta de homogeneidad	$h_3(C_m) = \min_{\nu=1\ldots,n_m} \left[\sum_{l=1}^{ml} (\delta_{ml,m\nu})^r\right]$
Separación	$i_1(C_m) = \sum_{l=1}^{n_m} \sum_{k\neq m} \sum_{\nu=1}^{n_k} (\delta_{ml,k\nu})^r$
Separación	$i_2(C_m m) = \min_{\substack{l=1\ldots,n_m \\ k\neq m \\ \nu=1,\ldots,n_k}} [(\delta_{ml,k\nu})^r]$

ejemplo:

$$c_1(n,g) = \sum_{m=1}^{g} h(m), \tag{1.28}$$

$$c_2(n,g) = \max_{m=1,\ldots,g}[h(m)] \tag{1.29}$$

o

$$c_3(n,g) = \min_{m=1,\ldots,g}[h(m)]. \tag{1.30}$$

El criterio en (1.29) determina la falta de homogeneidad dentro de los clústeres, mientras que c_2 y c_3 reflejan la falta de homogeneidad del 'mejor' y 'peor' grupo , respectivamente. Se observa que el criterio $\sum_{m=1}^{g} h_1(m)$ depende de los tamaños de los grupos, en consecuencia la suma podría ser muy grande, por lo tanto, un criterio más factible, en este caso, es $c_1^*(n,g) = \sum_{m=1}^{g} \frac{h_1(m)}{n_m}$.

Análogamente, se pueden definir criterios donde se consideren los índices de separación. Cuando se hace uso de los criterios de falta de homogeneidad se busca un clúster que minimice el criterio utilizado, en tanto que al trabajar con los índices de separación, se desea maximizarlos.

Si de datos continuos se trata, los criterios se derivan de una descomposición

de la matriz de dispersión, T [1], de dimensión $p \times p$, definida como

$$T = \sum_{m=1}^{g} \sum_{l=1}^{n_m} (x_{ml} - \overline{x})(x_{ml} - \overline{x})^T, \tag{1.31}$$

donde x_{ml} es el vector de dimensión p de las observaciones del objeto l en el grupo m y $\overline{x}$ es el vector de dimensión p de las medias de la muestra general para cada variable [1]. Dicha matriz es equivalente a la suma de las matrices W y B, $T = W + B$, con W la matriz de dispersión dentro de los grupos y B la matriz de dispersión entre grupos como en (1.3) [1].

En el caso particular cuando $p = 1$, W representa la suma de cuadrados dentro de los grupos y B la suma de cuadrados entre grupos, un criterio natural para este caso sería elegir la partición que minimice W o, de forma equivalente, que maximice B.

Sin embargo no es tan fácil cuando $p > 1$. En este caso, se puede recurrir a una extensión análoga al caso univariado de la minimización de la suma de cuadrados dentro de los grupos, aquí se desea minimizar las sumas de cuadrados dentro de los grupos pero ahora sobre todas las variables; en otras palabras, se desea minimizar la $tr(W)$. Se puede mostrar que esto es equivalente a minimizar la suma de las distancias euclidianas al cuadrado entre los individuos y la media de su grupo, o sea,

$$\sum_{m=1}^{g} \sum_{l=1}^{n_m} (x_{ml} - \overline{x}_m)(x_{ml} - \overline{x}_m)^T = \sum_{m=1}^{g} \sum_{l=1}^{n_m} d^2_{ml,m}, \tag{1.32}$$

con $d_{ml,m}$ la distancia Euclidiana entre el individuos l en el grupo m y la media del grupo m.

Además, este criterio también puede ser derivado de la matriz de disimilitudes de donde se puede mostrar que la minimización de $tr(W)$ también es equivalente a minimizar el criterio en (1.28) haciendo uso del índice $h_1(m)$ con $r = 2$ [1].

Una vez que se ha elegido el criterio que sugiere la mejor manera de agrupar a los objetos, la pregunta natural que surge es cómo encontrar una partición en g grupos que optimice dicho criterio. Inconscientemente, se podría calcular

el valor del criterio para cada partición y seleccionaría la que proporcione un valor óptimo para el criterio. No obstante, el número de particiones diferentes de n objetos en g grupos hace que esta tarea no sea sencilla, pues existe un gran número de particiones incluso cuando n y m son pequeños. Aún con el avance tecnológico actual, la cantidad de cálculos para trabajar sobre todas estas particiones es extremadamente grande y difícil de manejar. Por lo que surge de manera necesaria una serie de algoritmos conocidos como "hill climbing", también llamado algoritmo de escalada simple o ascenso de colinas para hallar el valor óptimo, éstos consisten en reorganizar las particiones existentes y mantener la nueva solo si ésta proporciona un mejor resultado. La secuencia general de pasos para desarrollar los algoritmos ya mencionados se encuentran en la Figura 1.4 :

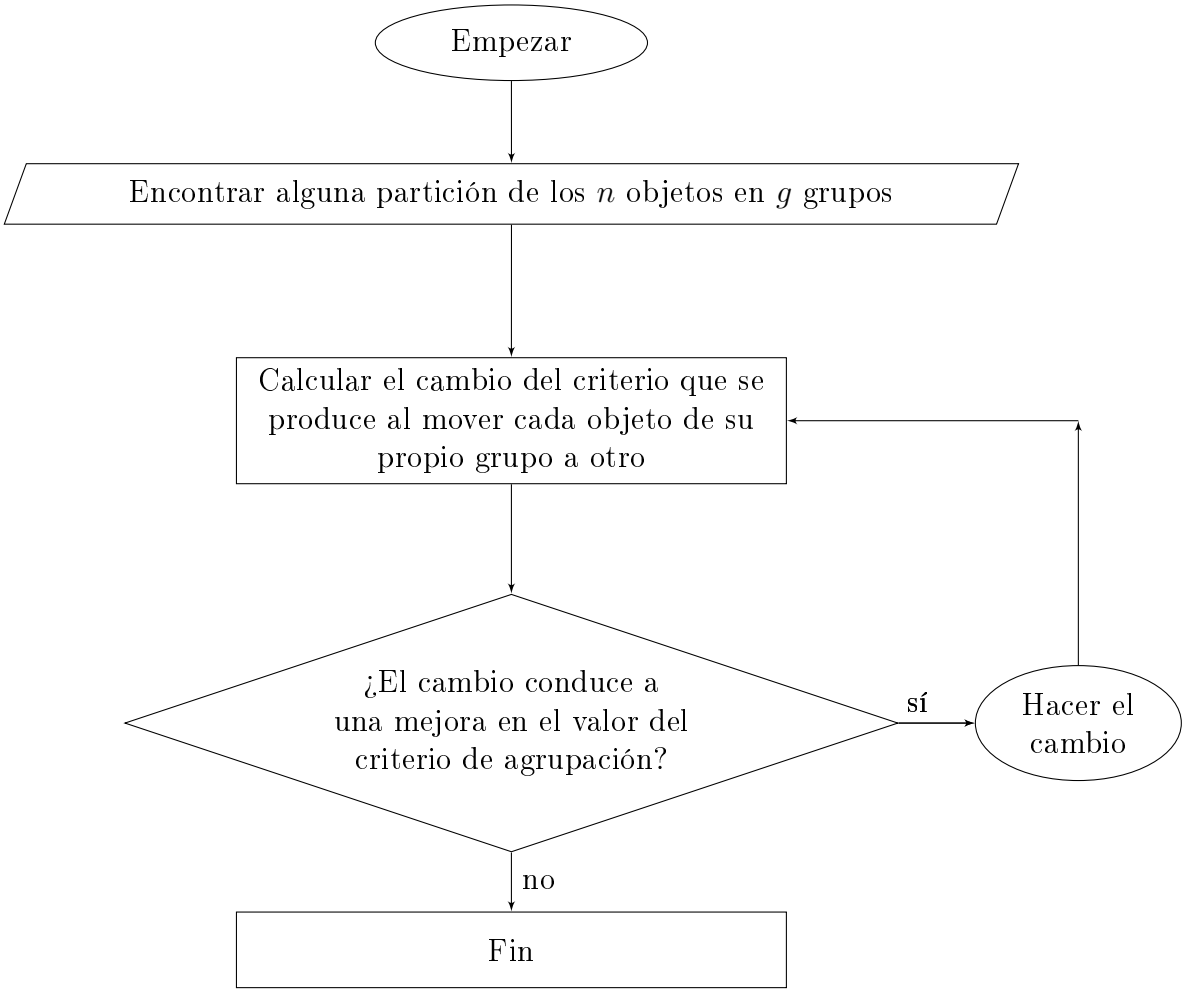


Figura 1.5: Diagrama de flujo del algoritmo tipo "hill climbing".

1. Encontrar alguna partición de los n objetos en g grupos.
2. Calcular el cambio que se produce en el criterio al mover cada objeto de su propio grupo a otro.
3. Realizar el cambio que conduzca a la mayor mejora del valor del criterio.
4. Repetir los dos pasos anteriores hasta que ningún movimiento de un solo objeto provoque que el criterio de agrupación mejore.

El problema que surge ahora es la manera de elegir una partición inicial. Podría, por ejemplo, elegirse basándose en conocimientos previos de los datos o ser

resultado de una aplicación previa de otro método, también se puede elegir una partición al azar. Pero los resultados se ven afectados por la elección de la partición inicial, esto es, diferentes particiones conducen a diferentes óptimos del criterio seleccionado [1], por lo que se recomienda ejecutar un algoritmo de optimización varias veces con particiones iniciales diferentes.

Uno de los primeros algoritmos de tipo "hill climbing.[en] plantearse consistía en actualizar iterativamente una partición reubicando simultáneamente cada objeto en el grupo a cuya medida está más cerca y luego recalculando las medias del grupo. Los algoritmos que involucran el cálculo de la media (centroide) de cada grupo, a menudo se denominan algoritmos de *k-means* [1].

El criterio de optimización que se utiliza en el algoritmo de *k-means* es la $tr(W)$. Este método es muy conocido y utilizado, busca la homogeneidad intra grupos, por lo que se desea minimizar dicho criterio. El diagrama presentado en la Figura 1.6 muestra la descripción general del algoritmo.

Como ya se ha mencionado, el objetivo es dividir la muestra en un número de grupos prefijado por el investigador, iniciando con los centros que representarán a cada grupo. El procedimiento que se debe seguir requiere de las cuatro etapas siguientes [4]:

1. Seleccionar k puntos como centroides de los grupos iniciales.

2. Calcular las distancias euclidianas de cada elemento al centro de los k grupos iniciales, y asignar cada elemento al grupo más próximo. La acción se realiza secuencialmente y al introducir un nuevo elemento en un grupo se recalculan los nuevos centroides.

3. Comprobar si reasignando uno a uno cada elemento de un grupo a otro se reduce $tr(W)$.

4. Si no es posible reducir $tr(W)$, terminar el proceso.

Es importante mencionar que el criterio de la traza tiene una propiedad muy importante: es invariante bajo cambios de medida en las variables. Cuando las

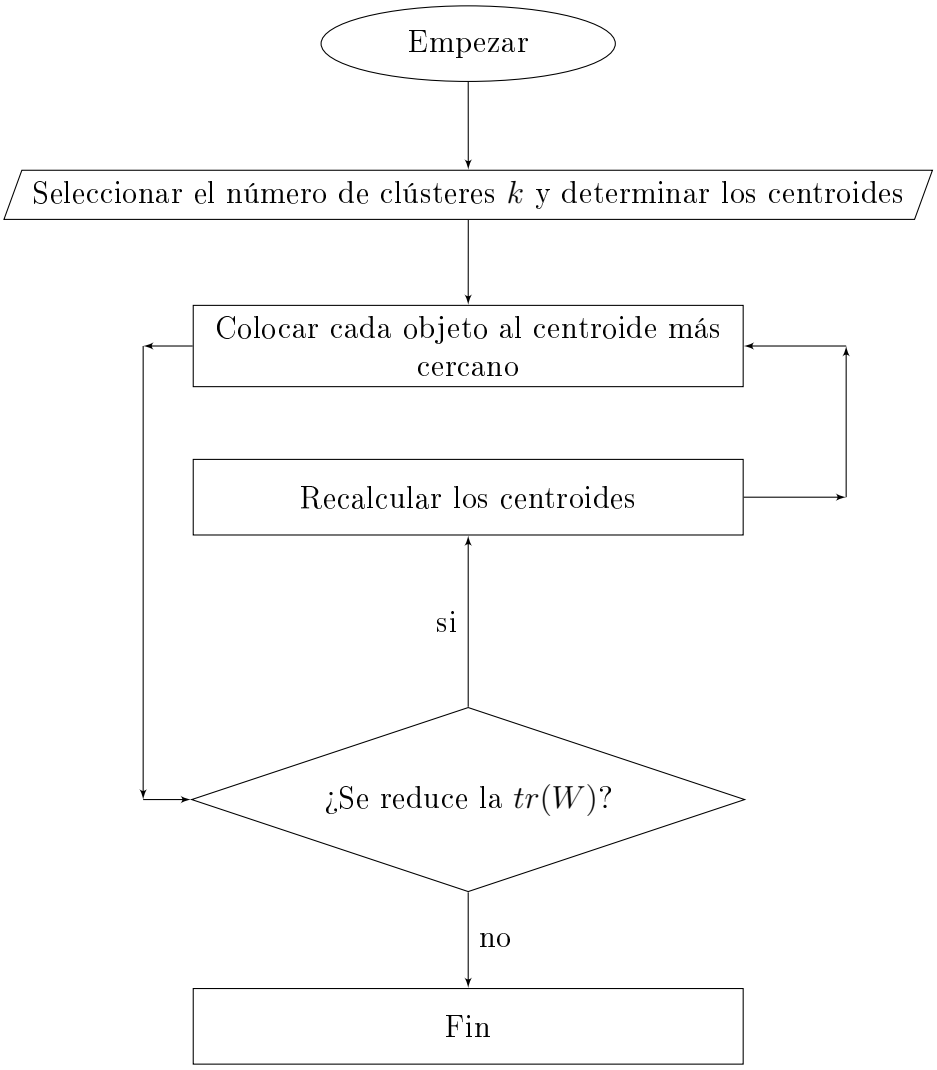


Figura 1.6: Diagrama de flujo del algoritmo k-means.

variables vayan en unidades distintas conviene estandarizarlas, para evitar que el resultado del algoritmo de *k-means* dependa de cambios irrelevantes en la escala de medida.

Por otro lado, *k-medoids* es una variación de *k-means*. En lugar de usar el punto medio como el centroide de un grupo, *k-medoids* usa un objeto del grupo para representarlo, *medoid* es el objeto ubicado más al centro del clúster, seleccionado como el de suma mínima de distancias a otros individuos. La idea básica del método es la siguiente: se seleccionan k puntos representativos para formar grupos iniciales (como en *k-means*). Y los objetos se mueven repetidamente al grupo que los represente mejor, según su medoid. Se analizan todas las combinaciones posibles de puntos representativos y no representativos y se calcula la calidad de la agrupación resultante para cada par. Un punto representativo original se reemplaza con el nuevo punto que produce la mayor reducción en la suma de distancias a otros objetos. En cada iteración, el conjunto de mejores puntos para cada grupo forma los nuevos medoids respectivos [5].

1.5. Implementación de técnicas de clústeres transversales a datos del sector turístico de México.

En esta sección se implementarán algunas de las técnicas descritas anteriormente, con ayuda del sofware **R** [26].

Para ello, se hace uso de una base de datos recuperada de la página del **INEGI** que reúne observaciones de los 32 estados de la República Mexicana, relacionadas con los ingresos por suministro de bienes y servicios e indicadores económicos del sector privado y paraestatal que tuvieron actividades relacionadas con el turismo en 2018, según identidad federativa. Tales observaciones se dividen en 3 variables: Número de establecimientos (unidades económicas), personal ocupado (personas

que trabajaban en las unidades económicas durante el periodo de referencia) y participación (porcentaje de las remuneraciones en los gastos totales de las unidades económicas).

Las funciones *hclust* del paquete **stats** [26] y *agnes* del paquete **cluster** [20] son las más populares que implementan métodos jerárquicos aglomerativos politéticos .

Para empezar se trabaja con la función *hclust* considerando los siguientes pasos:

1. Se exporta la base de datos de excel con la función *read_ excel*

```
Establecimientos<-read_excel("C:/Users/HP/Downloads/Arely1/Establecimi
entos.xlsx")
head(Establecimientos)
```

```
# A tibble: 6 × 4
  Estados             `Número de establecimientos` `Personal ocupado` Participación
  <chr>                                      <dbl>               <dbl>         <dbl>
1 Aguascalientes                                46                3166         14.8
2 Baja California                              166                7296         22.4
3 Baja California Sur                          193               32599          7.32
4 Campeche                                     112                2314         24.0
5 Coahuila de Zaragoza                         132                4183         22.6
6 Colima                                        74                4652         24.3
```

2. A continuación se eligen y escalan todas las variables numéricas:

```
Establecimientos1<-Establecimientos[ , c(2,3,4)]
Establecimientos<-scale(Establecimientos1)
```

3. Para la función *hclust* se requieren los valores de la distancia, que se pueden calcular en **R** utilizando la función *dist*. La medida de distancia predeterminada para la función *dist* es la euclidiana pero puede cambiarse con el argumento del método. También se necesita especificar el método de aglomeración que se desea usar, en este caso se hace uso del enlace completo.

```
d1<-dist(Establecimientos, method = "euclidean")
hc1<-hclust(d1, method="complete")
```

4. Se elige el número 'óptimo' de clústeres: En este caso se hace uso de un método muy popular en **R** que es *el método del codo*, el cual se basa en la observación de la inercia intra-clústeres. Sugiere que aumentar el número de clústeres puede ayudar a reducir dicha inercia. Para elegir el número 'adecuado' de clústeres se usa el punto de inflexión de la curva de la inercia con respecto al número de clústeres. La curva resultante mostrada en la Figura 1.7 sugiere que el valor 'más adecuado' de k es 5.

```
inertie <- sort(hc1$height, decreasing = TRUE)
plot(inertie[1:20], type = "l", xlab = "Nombre de classes", ylab = "
Inertie",lwd=2);grid()
k <- 5
abline(v=k,col="red",lty=5)
points(k,inertie[k],pch=16,cex=2,col="red")
```

   Puede verse que la inercia dentro de los grupos no parece variar demasiado después de $k = 5$. Aunque esta evaluación es subjetiva permite determinar un número aceptable de clústeres.

5. Luego, se pueden mostrar los clústeres obtenidos a través de un dendrograma como se muestra en la Figura 1.8:

```
plot(hc1, cex=0.6, hang=-1)
rect.hclust(hc1, k=4, border=2:10)
```

Dicha figura muestra que Quintana Roo se encuentra agrupado de manera individual. Colima, Jalisco, Oaxaca, Guerrero, Guanajuato, Veracruz, Estado de México, Ciudad de México y Puebla se encuentran en conjunto dentro de un mismo clúster. Aguascalientes, Sinaloa, Tabasco, Zacatecas, Campeche, Chihuhua, Chiapas, San Luis Potosí, Tamaulipas, Baja California Norte, Sonora, Coahuila, Querétaro y Yucatán coinciden dentro de otro clúster. Hidalgo, Durango, Tlaxcala, Michoacán y Morelos están categorizados dentro de un clúster diferente a los

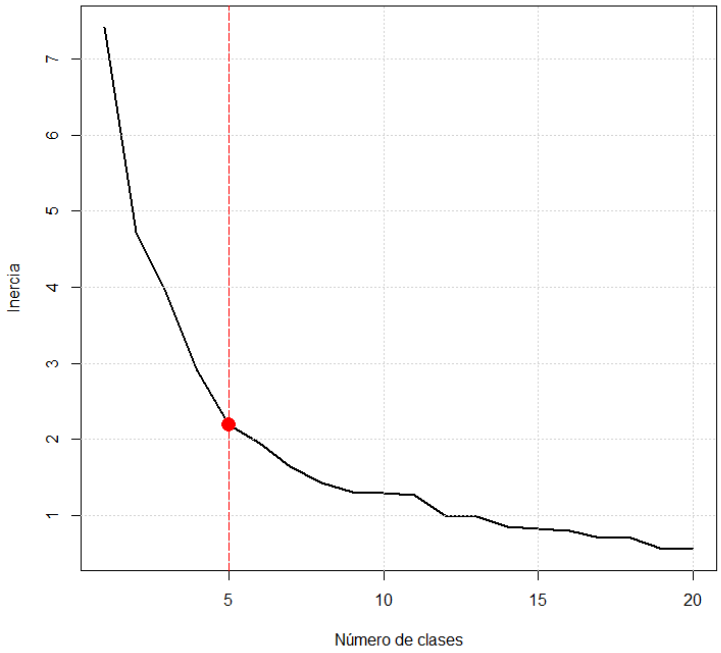


Figura 1.7: Gráfica de la inercia intraclústeres.

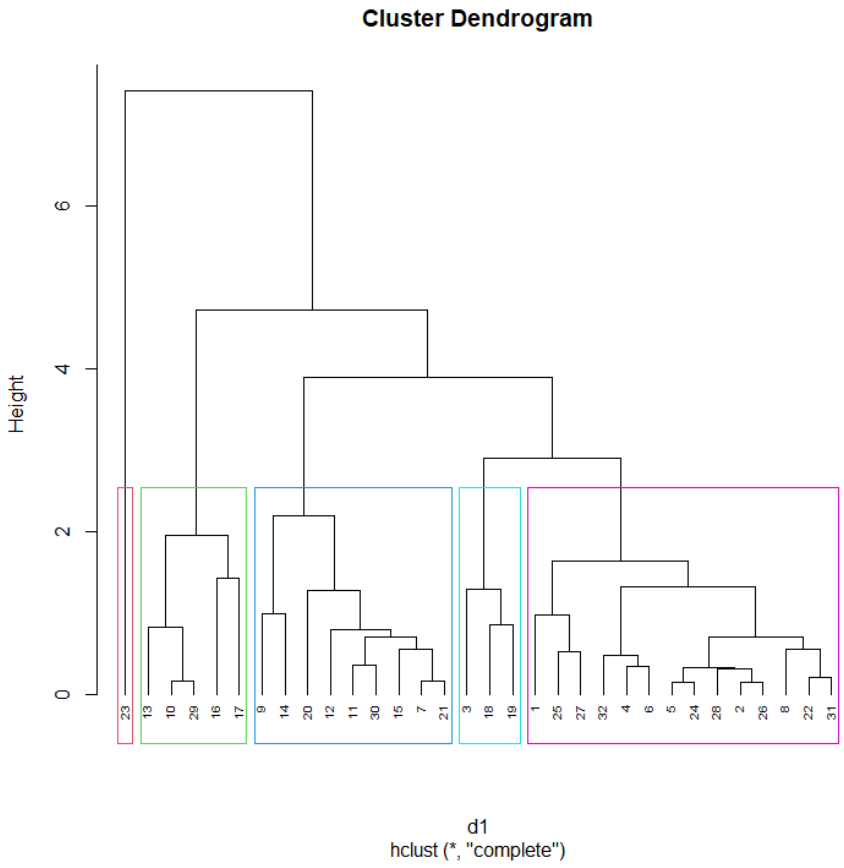


Figura 1.8: Clústeres obtenidos de la implementación de la función *hclust*.

anteriores. Mientras que, Baja California Sur, Nayarit y Nuevo León son elementos de un clúster distinto a los ya enumerados.

Ahora se puede aplicar el mismo razonamiento con la función *agnes*, considerando que el valor de k es el ya obtenido anteriormente:

```
agnes <- agnes(Establecimientos, method = "complete")

inertie <- sort(ag$height, decreasing = TRUE)
plot(inertie[1:20], type = "l", xlab = "Número de clases", ylab =
"Inercia",lwd=2);grid()
k <- 5
abline(v=k,col="red",lty=5)
points(k,inertie[k],pch=16,cex=2,col="red")
```

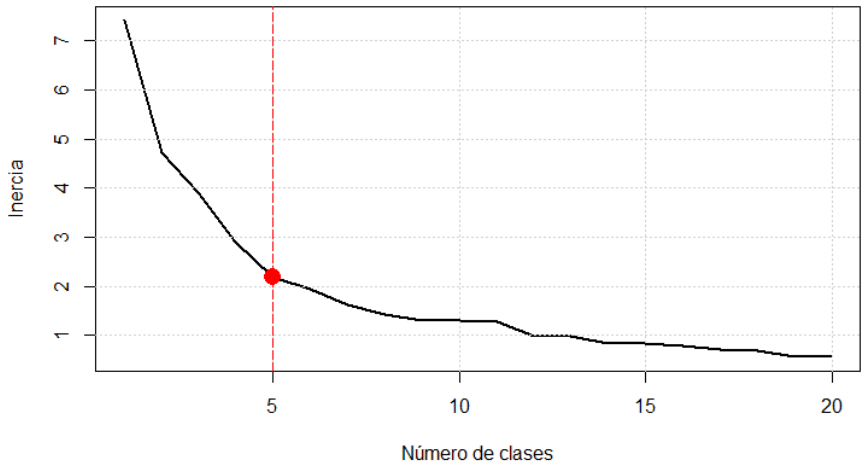


Figura 1.9: Gráfica de la inercia intraclústeres.

```
pltree(ag, cex = 0.6, hang = -1, main = "Dendrogram of agnes")
rect.hclust(ag, k = 5, border = 2:10)
```

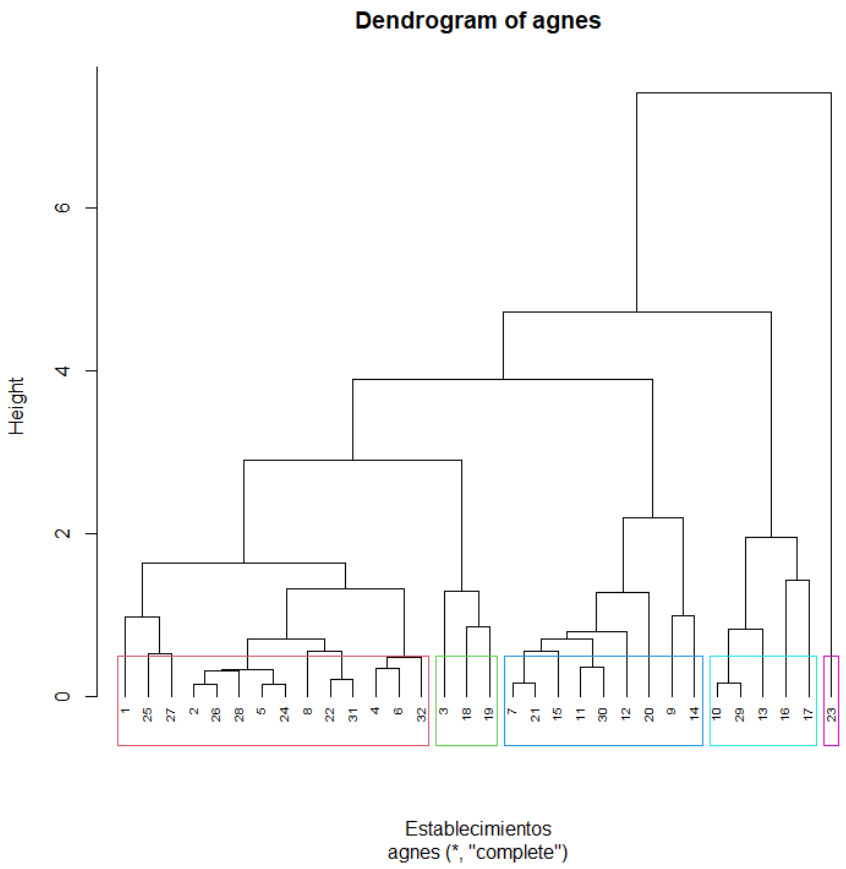


Figura 1.10: Clústeres resultantes de la implementación de la función *agnes*.

En este caso, como puede verse en la Figura 1.10, Quintana Roo también se encuentra agrupado de manera individual en un clúster. En otro, se encuentran Colima, Jalisco, Oaxaca, Guerrero, Guanajuato, Veracruz, Estado de México, Ciudad de México y Puebla. Aguascalientes, Nuevo León, Sinaloa, Tabasco, Zacatecas, Campeche, Chihuhua, Chiapas, San Luis Potosí, Tamaulipas, Baja California Norte, Sonora, Coahuila, Querétaro y Yucatán pertenecen a un clúster distinto a los anteriores. Hidalgo, Durango, Tlaxcala, Michoacán y Morelos pertenecen a otro clúster. Finalmente, en otro grupo están Baja California Sur y Nayarit.

Se puede observar que con las funciones *hclust* y *agnes* se obtiene sustancialmente la misma agrupación. Esto parece lógico sabiendo que ambos implementan métodos jerárquicos aglomerativos.

Por otro lado, también se presentan métodos divisivos. Comenzando por uno muy famoso que se implementa mediante la función *diana* del paquete *cluster*, considerando el mismo número de clústeres hallado anteriormente:

```
dian<-diana(Establecimientos)

pltree(dian, cex = 0.6, hang = -1, main = "Dendrogram of diana")
rect.hclust(dian, k = 5, border = 2:10)
```

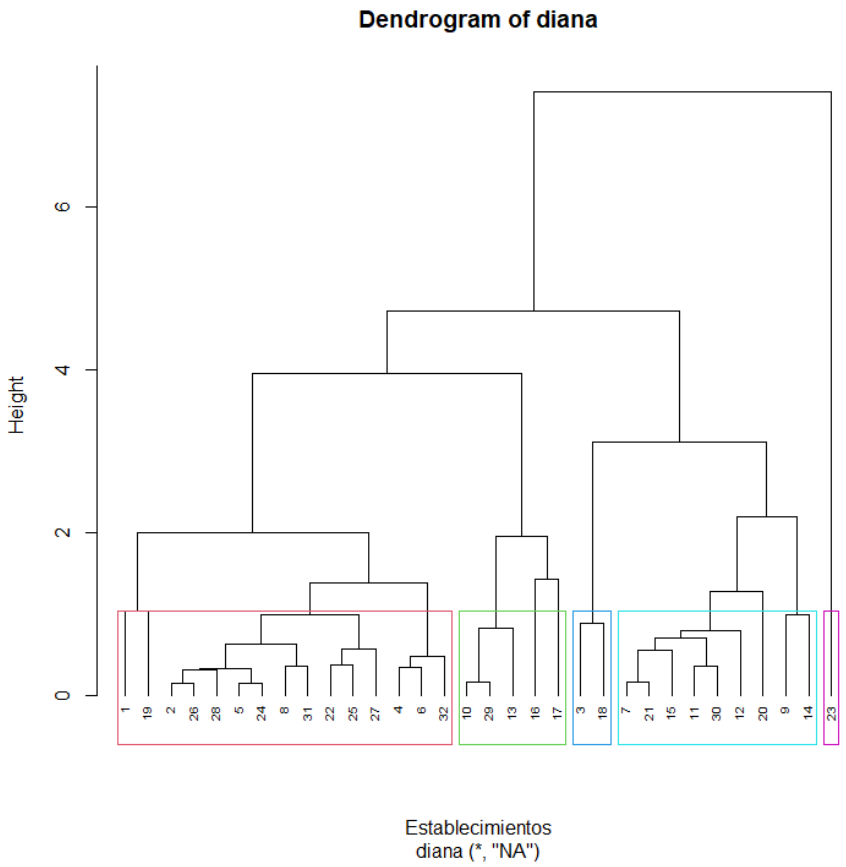


Figura 1.11: Clústeres resultantes de la implementación de la función *diana*.

Los resultados para este caso (Figura 1.11) indican que Aguascalientes, Nuevo León, Baja California Norte, Sonora, Tamaulipas, Chiapas, San Luis Potosí, Coahuila, Yucatán, Querétaro, Sinaloa, Tabasco, Campeche, Chihuahua y Zacatecas se encuentran agrupados dentro del mismo clúster. Ciudad de México,

Puebla, Estado de México, Guanajuato, Veracruz, Guerrero, Oaxaca, Colima y Jalisco están en otro clúster. En un clúster diferente a los demás, se agrupan Durango, Tlaxcala, Hidalgo, Michoacán y Morelos. Baja California Sur y Nayarit están dentro de otro clúster. Finalmente, Quintana Roo es el único elemento que pertenece a un clúster de forma individual, como pasó en los casos anteriores.

A continuación se presenta un método divisivo menos conocido, con ayuda de la función *divclust* del paquete **devtools** [12], siguiendo la siguiente secuencia de pasos:

1. Se crea el dendrograma que en este caso es un árbol de decisión, como se muestra a continuación (Figura 1.12):

```
tree<-divclust(Establecimientos)
plot(tree)
```

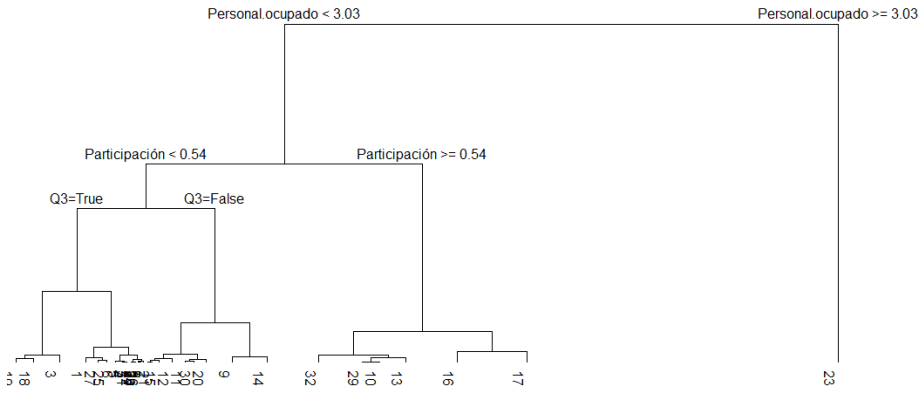


Figura 1.12: Árbol de decisión.

2. Se observa que, aplicando nuevamente el método del codo, la Gráfica 1.13 resultante sugiere (como en el caso anterior) que el número 'adecuado' de clústeres es 5:

```
plot(1:(tree$kmax-1),tree$height,xlab="number of cluster",ylab="height
",main="Split levels")
```

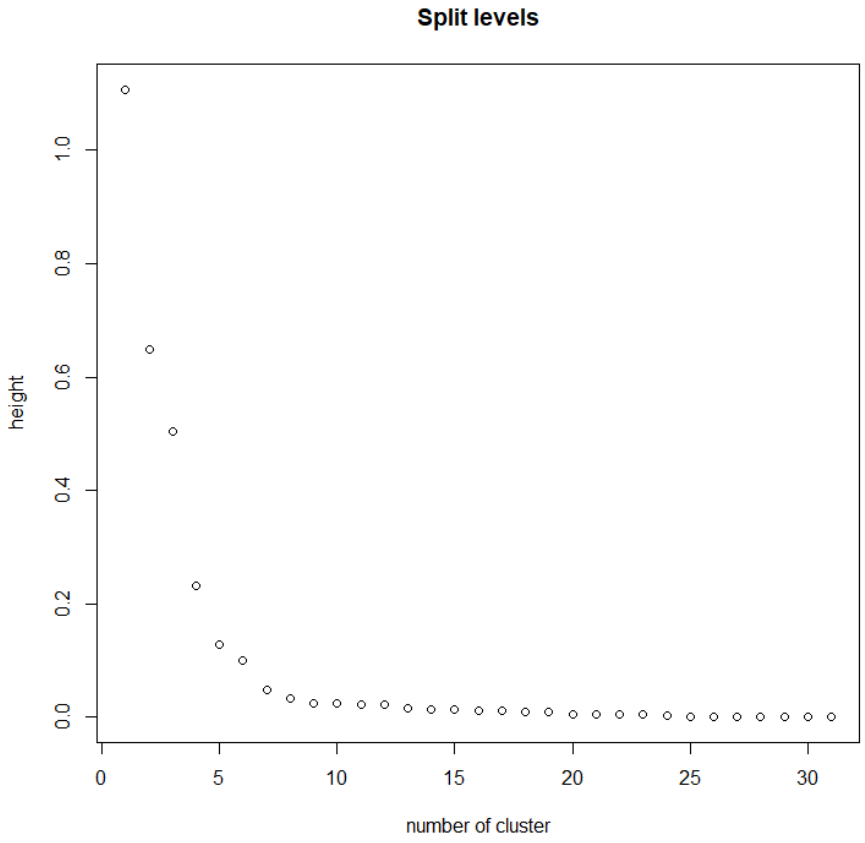


Figura 1.13: Gráfica de la inercia intraclústeres

3. Se considera el número de clústeres hallado anteriormente y se procede a graficar el árbol de decisión (1.14) con dicho número de clústeres, donde Q3: Número.de.establecimientos < 0.26 o Número.de.establecimientos ≥ 0.26 y Q4 : Participación < -1.03 o Participación ≥ -1.03.

```
c_5 <- divclust(Establecimientos, K=5)
plot(c_5,nqbin=4)
```

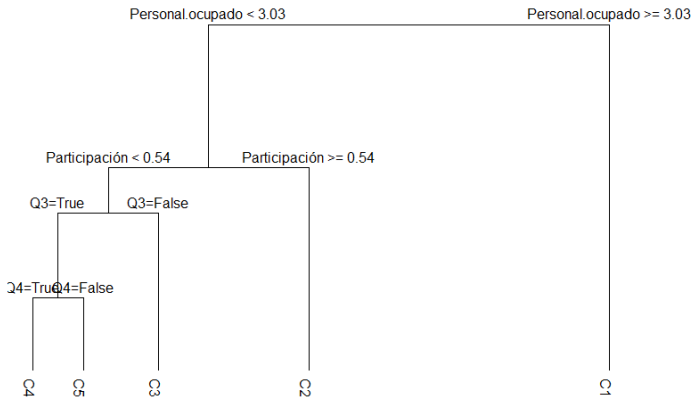


Figura 1.14: Árbol de decisión.

4. Luego, para conocer más acerca de la agrupación que sugiere *divclust,* se describe la inercia intraclústeres (Figura 1.15) y se recupera la lista de observaciones en cada uno de los clústeres:

```
c_5$B*100 #Inercia
c_5$clusters  # Recuperar la lista de observaciones en cada clústeres
```

```
> c_5$B*100 #explained inertia
[1] 83.00034
```

Figura 1.15: Inercia intraclústeres.

Las listas mostradas en la Figura 1.16 indican que nuevamente Quintana Roo se encuntra agrupado de forma única en un clúster. Que Durango, Hidalgo, Michoacán, Morelos, Tlaxcala y Zacatecas se agrupan en otro clúster. En un clúster diferente a los anteriores se encuentra Ciudad de México, Colima, Guanajuato, Guerrero, Jalisco, Oaxaca, Puebla y Veracruz. Aguascalientes, Baja California Norte, Campeche, Chiapas, Chihuahua, Coahuila, Queretaro, San Luis Potosí,

```
> c_5$clusters  # retrieve the list of observations in each cluster
$C1
[1] "23"

$C2
[1] "10" "13" "16" "17" "29" "32"

$C3
[1] "7"  "9"  "11" "12" "14" "15" "20" "21" "30"

$C4
[1] "3"  "18" "19"

$C5
 [1] "1"  "2"  "4"  "5"  "6"  "8"  "22" "24" "25" "26" "27" "28"
[13] "31"
```

Figura 1.16: Lista de observaciones en cada uno de los clústeres resultnates.

Sinaloa, Sonora, Tabasco, Tamaulipas, y Yucatán están dentro de otro clúster. Finalmente, Baja California Sur, Nayarit, y Nuevo León pertenecen a otro distinto.

Se observa que los clústeres resultantes con estos dos métodos jerárquicos divisivos son muy parecidos, como ocurrió con las agrupaciones resultantes en los métodos jerárquicos aglomerativos.

Finalmente, se implementará uno de los métodos no jerárquicos más conocido, *k-means*, el cual está basado en un criterio de optmización (Sección 1.4). Para implementar éste se consideran los siguientes pasos:

1. Como se sabe, este método requiere que se indique de antemano el número de clústeres que se desean crear, para determinarlo, como en las implementaciones anteriores, se hace uso de método del codo:

```
wssplot <- function(Establecimientos, nc=15, seed=1234){
  wss <- (nrow(Establecimientos)-1)*sum(apply(Establecimientos,2,var))
  for (i in 2:nc){
    set.seed(seed)
    wss[i] <- sum(kmeans(Establecimientos, centers=i)$withinss)}
  plot(1:nc, wss, type="b", xlab="Number of Clusters",
```

```
        ylab="Within groups sum of squares")}
wssplot(Establecimientos)
```

De lo que resulta, la gráfica de la Figura 1.17, en la cual se puede ver que a partir de 5 clústeres la inercia intra-clústeres se estabiliza, lo que indica que $k = 5$ es una 'buena' elección.

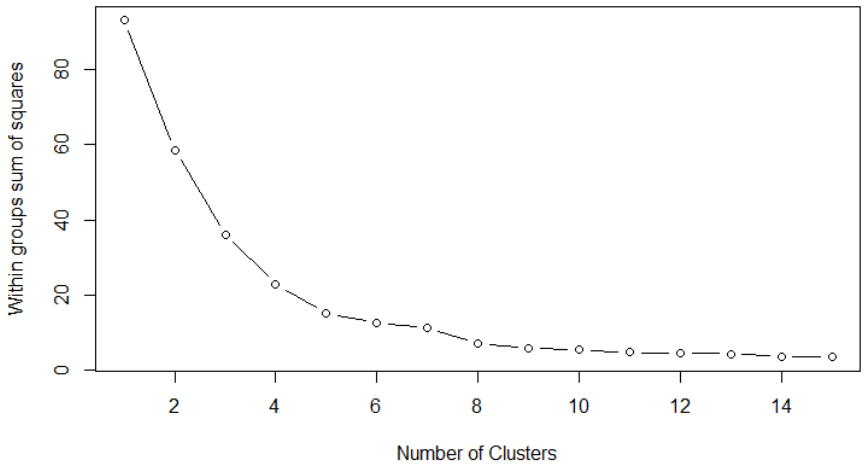


Figura 1.17: Inercia intraclústeres.

2. Se utiliza la función *k-means* para hallar el agrupamiento final:

```
set.seed(123)
kmeans <- kmeans(Establecimientos, 5, iter.max = 10000, nstart = 50)
```

3. Se extrae el tamaño de cada uno de los clústeres.

```
kmeans$size
```

```
> kmeans$size
[1]  9  1  5  3 14
```

4. Se extrae la asignación de las obvervaciones a los clústeres.

```
kmeans$clusters
```

```
> kmeans$cluster
 [1] 5 5 4 5 5 5 1 5 1 3 1 1 3 1 1 3 3 4 4 1 1 5 2 5 5 5 5 5 3 1 5 5
```

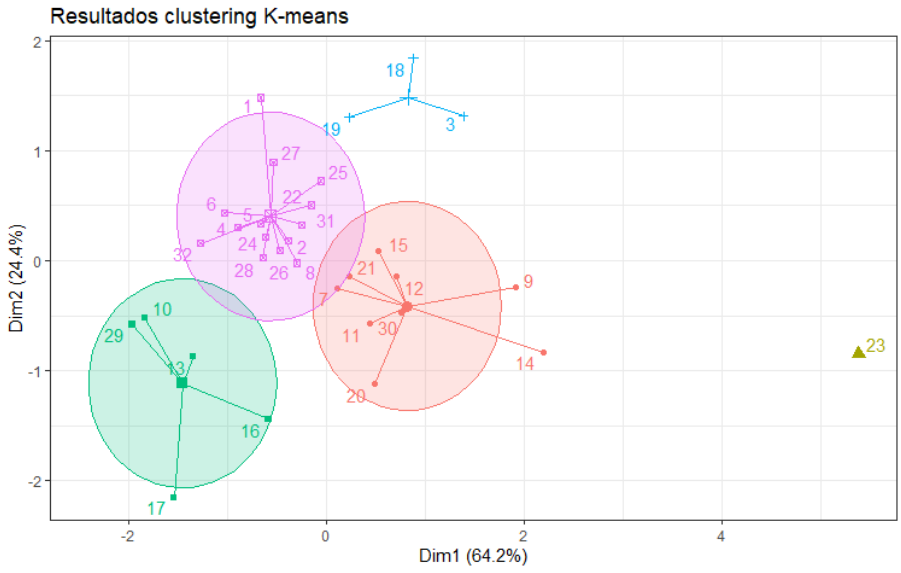
5. Se extrae el centro para cada clúster.

```
kmeans$centres
```

```
> kmeans$centers
  Número de establecimientos Personal ocupado Participación
1                  1.0797290        0.1063785    -0.2349234
2                  2.4662609        5.0008670    -1.4707780
3                 -0.5199145       -0.4235341     1.7668204
4                 -0.3357278        0.3306225    -1.6355500
5                 -0.6126476       -0.3451765    -0.0244545
```

6. Se obtienen las visualizaciones de las agrupaciones resultantes. Cuando el número de variables es mayor que 2 (como en este caso), automáticamente se representan las primeras dos variables.

```
fviz_cluster(object = kmeans, data = Establecimientos, show.clust.cent
 = TRUE,
             ellipse.type = "euclid", star.plot = TRUE, repel = TRUE) +
  labs(title = "Resultados clustering K-means") +
  theme_bw() +
  theme(legend.position = "none")
```



A partir de la implementación de $k-means$ se obtiene que dentro de un clúster se agrupan Aguascalientes, Baja California Norte, Campeche, Chiapas, Chihuahua, Coahuila, Querétaro, San Luis Potosí, Sinaloa, Sonora, Tabasco, Tamaulipas, Yucatán y Zacatecas. A un clúster distinto pertenecen Baja California Sur, Nayarit y Nuevo León. Mientras que, Durango, Hidalgo, Michoacán, Morelos y Tlaxcala están dentro de otro clúster. Ciudad de México, Colima, Guanajuato, Guerrero, Jalisco, Estado de México, Oaxaca, Puebla y Veracruz se encuentra en otro clúster diferente. Finalmente, Quinatana Roo, como en los casos anteriores, también se encuentra agrupado de forma individual en un cluster distinto a los ya enumerados previamente.

Capítulo 2

Técnicas de clústeres para datos longitudinales

La agrupación de datos longitudinales es de suma importancia debido a que se ha convertido en una herramienta escencial para realizar investigaciones médicas, epidemológicas, biológicas y sociales; entre las que destacan la clasificación de pacientes, de clientes, de organismos, étc., que evolucionan en distintos aspectos a lo largo del tiempo [6].

La característica definitoria de un estudio longitudinal es que los individuos se miden repetidamente a través del tiempo. Los estudios longitudinales contrastan con los estudios transversales en que los de tipo transversal miden un resultado único para cada individuo. Si bien, frecuentemente es posible abordar los mismos aspectos científicos a través estudio longitudinal o transversal, la principal ventaja del primero es su capacidad para separar lo que en el contexto de estudios de población se denomina efectos de cohorte y periodo (Diggle, 2002)[24].

En el caso de datos longitudinales, la sucesión de observaciones se nombra comúnmente como trayectorias o series de tiempo, las cuales pueden ser simples o conjuntas. Igual que en las técnicas transversales, se considera un conjunto de datos compuesto por n objetos descritos por p variables X_A, X_B, $\cdots$, X_P, pero medidas en t instantes distintos, donde a X_A se llama variable trayectoria y a

la colección de variables trayectorias $(X_A, X_B, \cdots, X_P)$ se le denomina variable trayectoria cojunta. Luego, para el sujeto i, el valor de X_A en el momento j se denota como x_{ijA} y a $x_{i.A} = (x_{i1A}, x_{i2A}, \cdots, x_{itA})$ se le llama trayectoria simple de la variable A para el individuo i. Por otra parte, se le denomina trayectoria conjunta a la matriz de dimensión $p \times t$ cuyas filas son las trayectorias simples y las columnas corresponden al estado del individuo i en el instante j

$$x_i := \begin{pmatrix} x_{i1A} & x_{i2A} & \cdots & x_{itA} \\ x_{i1B} & x_{i2B} & \cdots & x_{itB} \\ \vdots & \vdots & \vdots & \vdots \\ x_{i1P} & x_{i2P} & \cdots & x_{itP} \end{pmatrix}.$$

Calcular distancias es una clave importante para la implementación de cualquier técnica para crear clústeres, el caso longitudinal no es la excepción. En este caso, hallar la distancia entre $x_{i..}$ y $x_{k..}$ plantea el problema de hallar distancias entre trayectorias, es decir, distancias entre matrices. La literatura se centra principalmente en dos métodos [2]. El primero consiste en:

1. Para cada $j = 1, \ldots, t$, calcular la distancia $d_{j.}(x_{ij.}, x_{kj.}) = Dist(x_{ij.}, x_{kj.})$ entre los pares de columnas, de donde resulta un vector de distancias

$$(d_{1.}(x_{i1.}, x_{k1.}), d_{2.}(x_{i2.}, x_{k2.}), \ldots, d_{t.}(x_{it.}, x_{kt.}))\,.$$

2. Usar una función que combine las t distancias calculadas

$$d(x_{i..}, x_{k..}) = \| \, (d_{1.}(x_{i1.}, x_{k1}), d_{2.}(x_{i2.}, x_{k2}), \cdots, d_{t.}(x_{it.}, x_{kt})) \, \| \,. \tag{2.1}$$

Donde $||\cdot||$ representa una distancia, en este caso, distancia entre columnas.

El segundo consiste en reproducir el mismo proceso, pero considerando las distancias entre filas en lugar de por columnas.

Las técnicas de clústeres para datos longitudinales tienen distintos enfoques [6]. En este trabajo se abordará solamente el enfoque por agrupamiento transversal. El cuál consiste en desplegar las técnicas propias del análisis transversal considerando a cada trayectoria conjunta (o simple) como la medición asociada con un individuo.

Esto implica que las matrices de distancias correspondientes están basadas en métricas definidas en espacios de matrices o vectores dependiendo de la naturaleza de las trayectorias. Entre los métodos trasversales más usados se encuentran los métodos longitudinales (aglomerativos y divisivos) y *k-means* [6].

2.1. K-means

En el contexto longitudinal, la técnica *k-means* , normalmente, se denomina *k-means longitudinal*. Es necesario que las observaciones tengan la misma longitud, esto significa que todos los individuos deben ser medidos en la misma cantidad de instantes.

Invariable al caso transversal, *k-means longitudinal* tiene como objetivo encontrar la partición que minimice la varianza dentro de los clústeres, siguiendo la misma lista de pasos que en el caso transversal. Pero, en este escenario, los centroides son trayectorias representativas de cada uno de los grupos y los grupos resultantes contendrán individuos que sigan dichas trayectorias.

kml y *kml3d* [3] son librerías de **R** creadas para implementar *k-means longitudinal*, el primero está diseñado para trabajar con trayectorias simples y el segundo con trayectorias conjuntas.

Al igual que en las técnicas usadas para datos transversales, es de suma importancia determinar cuál es el mejor número de clústeres, en la Sección (1.3) ya se han expuesto algunos criterios que ayudan a resolver esta cuestión. Los críterios que ofrecen estos paquetes son [2]:

1. Criterio de Calinski y Harabatz.

2. Criterio de Calinski y Harabatz, variante Kryszczuk.

$$C_G(k) = \frac{tr(B)}{tr(W)} \cdot \frac{n-k}{k-1}. \tag{2.2}$$

3. Criterio de Calinski y Harabatz, variante Genolini.

$$C_G(k) = \frac{tr(B)}{tr(W)} \cdot \frac{n-k}{\sqrt{k-1}}. \tag{2.3}$$

4. Criterio de Ray y Turi.

$$R(k) = \frac{V_{intra}}{V_{inter}}, \tag{2.4}$$

donde $V_{intra} = \sum_x (dist(x, centro(x))$ y $V_{inter} = min_{i,j}(dist(centro_i, centro_j)^2)$.

5. Criterio de Davies y Bouldin.

$$D(k) = media(P(C_i, C_j)), \tag{2.5}$$

donde

$$P(i,j) = \frac{DisInterna(C_i) + DistInterna(C_j)}{DistExterna(C_i, C_j)}, \tag{2.6}$$

siendo C_i, C_j los clústeres $i \neq j$, DistInterna es una medida de compacidad del clúster correspondiente (por ejemplo, distancia máxima entre dos puntos del clúster) y DistExterna es una medida de separación de clústeres (para ejemplo, distancia entre los centros de los clústeres).

Considerando que n es el número de individuos y $N = n \cdot t$:

6. $BIC = 2 \times log(L) - h \times log(n)$.
7. $BIC2 = 2 \times log(L) - h \times log(N)$.
8. $AIC = 2 \times log(L) - h \times h$.
9. $AICc = AIC + \frac{2h(h+1)}{n-h-1}$.
10. $AICc2 = AIC + \frac{2h(h+1)}{N-h-1}$.

El problema de la diferencia de escalas en las variables involucradas en este tipo de estudios, ocurre muy comúmente. La diferencia con los datos que no son de tipo longitudinal es que las trayectorias se estandarizan en conjunto y no en

cada instante [2]. Por ejemplo, si se denota a la media de $x_{..A}$ como $\overline{x_{..A}}$ y a la desviación típica como s_A, entonces la estandarización de x_{ijA} será

$$x_{ijA} = \frac{x_{ijA} - \overline{x_{..A}}}{s_A}, \tag{2.7}$$

y para estandarizar la trayectoria conjunta $x_{i..}$ se debe estandarizar cada una de las trayectorias simples por separado. Por otro lado, como ya se ha visto antes, el algoritmo de *k-means* inicia con la elección de un conjunto de k centroides. Se han propuesto métodos para esta inicialización y la mayoría de ellos están basados en construir una configuración inicial en la que los centroides estén posiblemente lo más elejados posible entre ellos. Los paquetes *kml* y *kml3d* ofrecen distintos métodos para elegir las configuraciones iniciales, los cuales se describen, de forma general, a continuación [2]:

1. **randomK**: Elige k individuos al azar.
2. **randomAll**: Asigna aleatoriamente a todos los individuos a un clúster y calcula el centroide de cada clúster como su media.
3. **maxDist**:
    a. Calcula matriz de distancia entre todos los puntos.
    b. Elige los dos puntos más alejados como los dos primeros centroides C_1 y C_2.
    c. Inicia un bucle:
        i. Para cada individuo x, considera $D(x)$, la distancia entre x y el centroide más cercano.
        ii. Elige como nuevo centroide C_i al individuo para el cual $D(C_i)$ es máximo.
        iii. Repite los pasos $3(c)i$ y $3(c)ii$ hasta que se hayan elegido k centros.

4. **kmeans+**: Se basa en el principio de **maxDist**, la única diferencia es que el primer centro se elige al azar, evitando el cálculo de la matriz de distancias y reduciendo el costo computacional.

5. **kmeans-**: Debido a que en **kmeans+** el primer punto seleccionado puede ser una "mala" elección. **kmeans-** trata de mejorar el método anterior, para ello selecciona aleatoriamente un centro y calcula la distancia a éste del resto de los puntos, luego selecciona como segundo centro aquel que está más lejos del primero, con lo que se asegura que este segundo centro es adecuado, pues es el más distante al menos de un punto. Luego, se lleva a cabo el siguiente bucle:

    i. Se considera la distancia de cada punto al centro más cercano.

    ii. Se elige como un nuevo centroide aquel punto cuya distancia es la mayor.

6. **kmeans- -**: Trata de mejorar a **kmeans-**, eligiendo aleatoriamente los centros que se van agregando a la lista, para que la probabilidad de que los centroides sean distantes unos de otros y la probabilidad de que un individuo seleccionado como centroide sea proporcional al cuadrado entre el individuo y los centros previamente seleccionados.

7. **kmeans++**:Es la versión no determinista de **kmeans+**.

Como ya se mencionó, *kml* y *kml3d* [3] en el software **R**, ejecutan el algoritmo *k-means*, ambos lo ejecutan variando los métodos de inicialización y el número de clústeres. Es importante mencionar que se brinda la opción de especificar el número de clústeres para los que se harán las particiones y el número de veces que se ejecuta el algoritmo; en caso de no especificarse, se ejecutan con los valores por defecto. Dichos valores son 2, 3, 4, y 6 clústeres; y 20 iteraciones.

Capítulo 3

Implemantación de k-means longitudinal en datos del sector turístico en México

El análisis de la realidad turística requiere muchas veces del estudio simultáneo de dos o más variables, con el objetivo de comprobar la relación entre ellas. En esta sección se hará uso del agoritmo k-means para encontarar agrupaciones de los 32 estados de la República Méxicana en función de la evolución del turismo. Para ello se hace uso de las librerías *kml* y *kml3d* con los datos de inicialización y el número de clústeres que se ofrecen por defecto.

Para empezar, se trabajó con las trayectorias medidas sobre el periodo de enero de 1992 a diciembre de 2019. Primero se hizo el análisis de las trayectorías individuales.

Variable: Total de visitantes

En primer lugar, se determinó el mejor número de clústeres a partir de los criterios de calidad, *kml* cuenta con la opción de visualizar varios criterios en una sola gráfica, que permite compararlos. Para que esto sea posible, se encuentran los valores contrarios en los criterios que deben minimizarse, de tal forma que

se pueda elegir el número de clústeres basados en la maximización de cada uno de estos. Además, los muestra estandarizados entre 0 y 1. En la Figura 3.1 se muestran los criterios que se tomaron en cuenta para elegir el número de clústeres óptimo. Puede observarse que tres de ellos concuerdan con la elección de cinco clústeres, por lo que se decidió realizar el análisis para ese número.

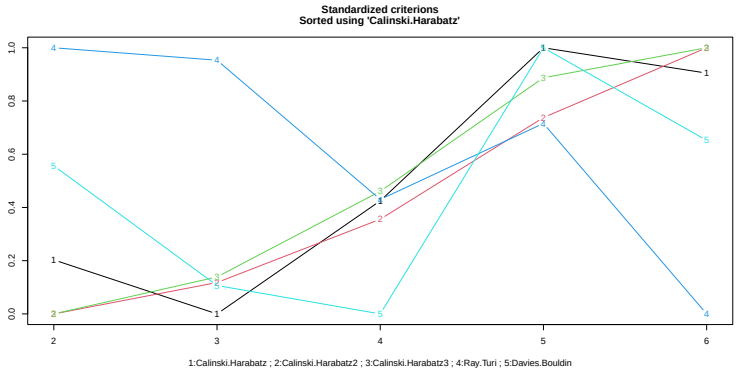


Figura 3.1: Criterios de calidad para determinar el número de clústeres.

Se obtiene entonces los siguientes agrupamientos basados en *kml* de **R:**

1. A: Aguscalientes, Baja California Sur, Campeche, Coahuila, Colima, Durango, Hidalgo, Morelos, Nayarit, Queretaro, San Luis Potosí, Tabasco, Tlaxcala, Yucatán y Zacatecas.

2. B: Baja California Norte, Chiapas, Chihuahua, Guanajuato, Estado de México, Michoacán, Nuevo León, Oaxaca, Puebla, Sinaloa, Sonora, Tamaulipas.

3. C: Guerrero, Jalisco y Veracruz.

4. D: Ciudad de México.

5. E: Quintana Roo.

En la Figura 3.2 se muestran las trayectorias de los centroides de cada uno de los grupos y las trayectorias pertenecientes a cada uno de ellos, las cuales están representadas del mismo color respectivamente.

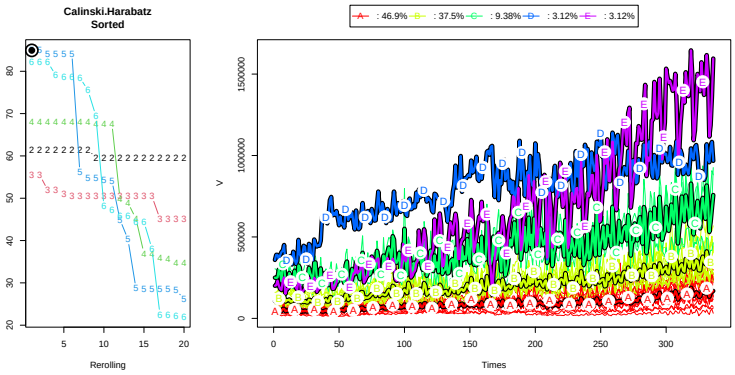


Figura 3.2: Representación de la agrupación para cinco clústeres considerando la variable "Total de visitantes".

Grupo	Porcentaje	Absoluto
A	46.9 %	15
B	37.5 %	12
C	9.38 %	3
D	3.12 %	1
E	3.12 %	1

Tabla 3.1: Distribución de las trayectorias para la variable "Total de visitantes."

Los resultados delatan la gran excedencia del número de visitantes de la Ciudad de México y Quintana Roo sobre los demás estados. Lo que remarca la veracidad de que Ciudad de México es considerado como el principal centro de actividades turísticas como consecuencia de ser una atractiva ciudad colonial, su pasado arqueológico y ser la sede del gobierno federal [22]. Por otro lado, en [8], basados en la Tasa de Intensidad turística (TIT), Tasa de penetración turística (TPT), Tasa de Densidad Turística (TDT) y Grado de Internacionalización Turística (GIT), mencionan que los centros turísticos pueden tipificarse de acuerdo a seis niveles, de los cuales solo Cancún se encuentra en la mayor de las jerarquías, lo que coincide con que también Quintana Roo se encuentre clusterizado de manera individual en un grupo.

Variable: Total de habitaciones disponibles

Procediendo de la misma manera que en la variable anterior, se puede ver en la Figura 3.3 que se puede elegir a seis como el número óptimo de clústeres. Obteniendo así la siguiente agrupación:

1. A: Baja California Norte, Baja California Sur, Chiapas, Chihuahua, Guanajuato, Estado de México, Michoacán, Nayarit, Nuevo León, Oaxaca, Puebla, Sinaloa, Sonora, Tamaulipas.

2. B: Aguascalientes, Campeche, Coahuila, Colima, Durango, Hidalgo, Morelos, Querétaro, San Luis Potosí, Tabasco, Tlaxcala, Yucatán y Zacatecas.

3. C: Guerrero, Veracruz.

4. D: Ciudad de México.

5. E: Jalisco.

6. F: Quintana Roo.

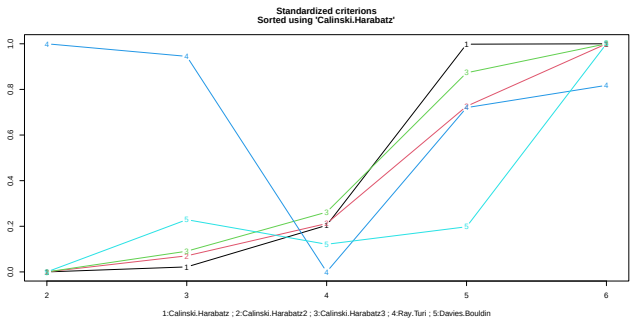


Figura 3.3: Criterios de calidad para determinar el número de clústeres.

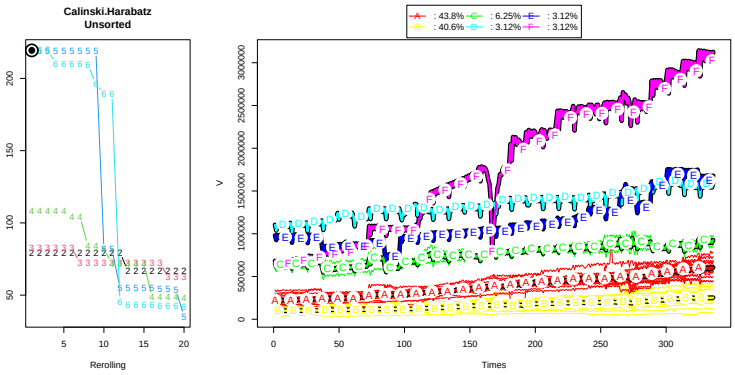


Figura 3.4: Representación de la agrupación para seis clústeres considerando la variable "Total de habitaciones disponibles".

Grupo	Porcentaje	Absoluto
A	43.8 %	14
B	40.6 %	13
C	6.25 %	2
D	3.12 %	1
E	3.12 %	1
F	3.12 %	1

Tabla 3.2: Distribución de las trayectorias para la variale "Total de habitaciones disponibles".

Variable: Porcentaje de ocupación

En este caso, según la Figura 3.5, puede considerarse hacer el análisis para seis clústeres.

Obteniendo como resultado los siguientes clústeres:

1. A:Campeche, Chihuahua, Colima, Jalisco, Puebla, Querétaro, San Luis Potosí, Sinaloa, Sonora, Tabasco y Yucatán.

2. B: Aguascalientes, Baja California Norte, Coahuila, Durango, Hidalgo, Tamaulipas, Veracruz y Zacatecas.

3. C: Chiapas, Guanajuato, Veracruz, Estado de México, Michoacán, Morelos, Oaxaca y Tlaxcala.

4. D: Baja California Sur, Ciudad de México y Nuevo León.

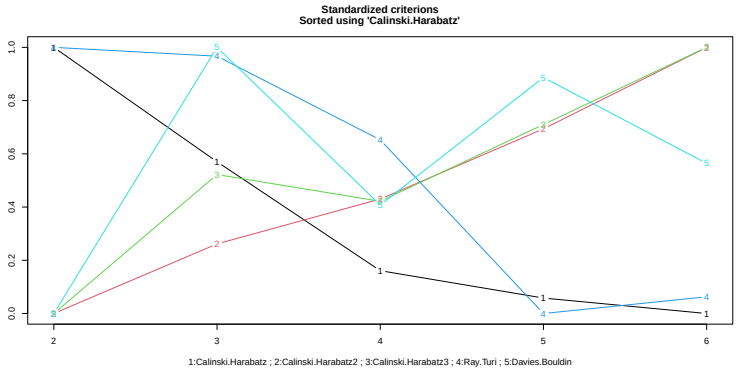


Figura 3.5: Criterios de calidad para determinar el número de clústeres.

5. E: Nayarit.

6. F: Quintana Roo.

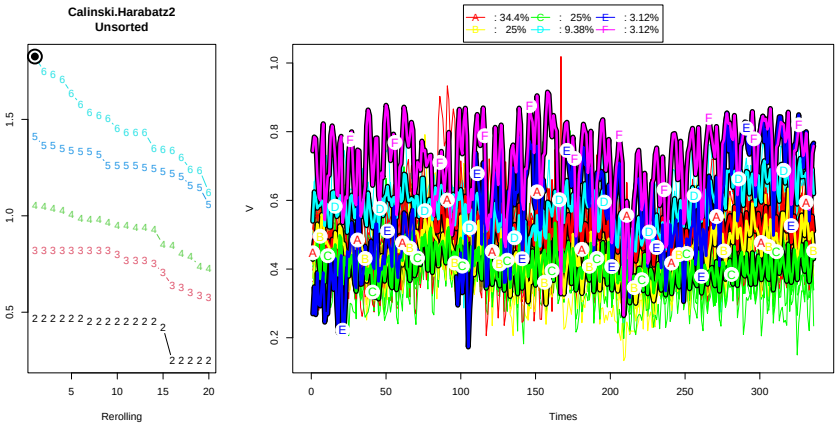


Figura 3.6: Representación de la agrupación para seis clústeres considerando la variable "Porcentaje de ocupación".

Grupo	Porcentaje	Absoluto
A	34.4 %	11
B	25 %	8
C	25 %	8
D	9.38 %	3
E	3.12 %	1
F	3.12 %	1

Tabla 3.3: Distribución de las trayectorias para la variale "Porcentaje de ocupación".

Tres variables Conjuntas

Después, con el fin de comparar los resultados arrojados por *kml* y *kml3*, se hizo el análisis de las trayectorias conjuntas. Como puede verse en la Figura 3.7 el criterio de Calinski y Harabatz sugirió que el mejor número de clústeres es 2, con lo que resulta la siguiente agrupación:

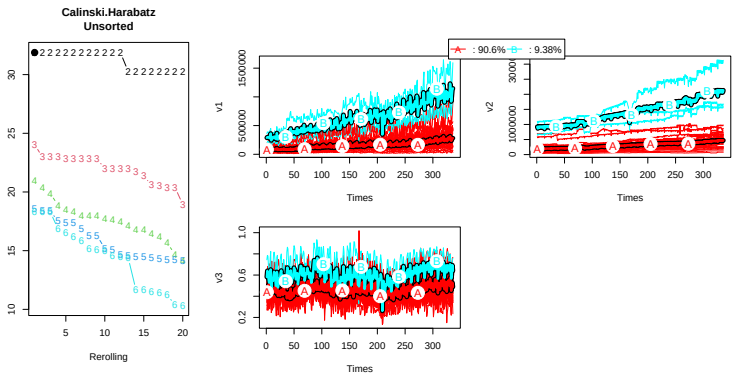


Figura 3.7: Representación de la agrupación para tres clústeres considerando las tres variables conjuntas y a los 32 estados.

1. A: Aguascalientes, Baja California Norte, Campeche, Chiapas, Chihuahua, Coahuila, Colima, Durango, Guanajuato, Guerrero, Hidalgo, Estado de México, Michoacán, Morelos, Oaxaca, Puebla, Querétaro, San Luis Potosí, Sinaloa, Sonora, Tabasco, Tamaulipas, Tlaxcala, Veracruz, Yucatán, Zacatecas, Baja California Sur, Nayarit y Nuevo León.

2. B: Ciudad de México, Jalisco y Quintana Roo.

Pero, los resultados al elegir 2 clústeres no brindan información basta acerca de los estados pertenecientes al grupo A, por lo que se decidió hacer una recluste-rización dentro del mismo. Para esto, se observó que esta vez el criterio de Calinski y Harabatz recomienda crear 3 clústeres (Figura 3.8).

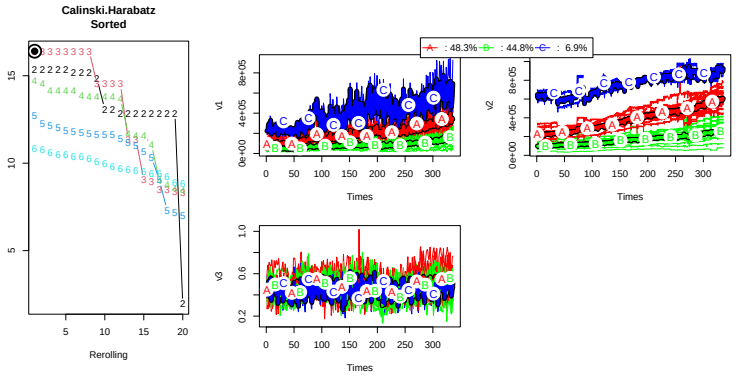


Figura 3.8: Representación de la agrupación para tres clústeres considerando las tres variables conjuntas y a los 32 estados.

Resultando finalmente una agrupación con 4 clústeres.

1. A. Baja California Norte, Baja California Sur, Chiapas, Chihuahua, Guanajuato, México, Michoacán, Nayarit, Nuevo León, Oaxaca, Puebla, Sinaloa, Sonora, Tamaulipas.

2. B: Aguascalientes, Campeche, Coahuila, Colima, Durango, Hidalgo, Morelos, Querétaro, San Luis Potosí, Tabasco, Tlaxcala, Yucatán y Zacatecas.

3. C: Guerrero y Veracruz.

4. D: Ciudad de México, Jalisco, Quintana Roo.

Además de las agrupaciones de los estados basados en el análisis del periodo de enero de 1992 a diciembre de 2019 usando kml y kml3d, se hizo el análisis para conseguir agrupaciones en cuatro subperiodos de tiempo (correspondientes a los sexenios presidenciales), también tomando en cuenta los casos donde se toman las trayectorias cojuntas y trayectorias simples. Esto, con el propósito de analizar

e identificar cómo cambia cada estado según el tiempo y la(s) variable(s) que se estén considerando.

A continuación se presentan los resultados correspondiente a los cuatro subperiodos, para las trayectorias simples y las trayectorias conjuntas: La elección del número de clústeres para cada caso se hizo de la misma forma que se procedió en los casos expuestos ateriormente.

Análisis de trayectorias simples para los cuatro subperiodos

Se realiza una clusterización al analizar cada una de las variables de forma aislada; es decir, se obtendrán tres agrupaciones (una para cada variable). Los resultados de este enfoque se observan en las Figuras 3.9, 3.10, 3.11, 3.12 y 3.13. En tales figuras se observan mapas en donde, estados con idéntico color pertenecen al mismo clúster al considerar la variable "Total de visitantes", estados con el mismo tramado pertenecen al mismo clúster al tomar en cuenta a la variable "Total de habitaciones disponibles" y el mismo color de frontera indica la pertenencia al mismo clúster para la variable "Porcentaje de ocupación". Cada mapa corresponde al periodo indicado en su correspondiente leyenda.

En la Figura 3.9 correspondiente al Sexenio 1, se observa que el número de clústeres a partir de las variables "Total de visitantes" y "Total de habitaciones disponibles" coincide, pero respecto a la variable "Porcentaje de ocupación" difiere. Además, puede notarse que la Ciudad de México resalta por ser el mejor estado con actividad turística considerando las variables ya mencionadas. Y Quintana Roo, Jalisco, Guerrero y Veracruz sobresalen como los estados con mayor cantidad de visitantes y número de habitaciones disponibles. Mientras que, Coahuila, Colima, Nayarit, Queretaro y Yucatán son los que menos destacan en esta actividad.

La Figura 3.10, muestra que igual que en el primer subperiodo la cantidad de clústeres coincide con respecto de las dos primeras variables, pero respecto a la tercera el número es menor. En este caso, el estado de Tamaulipas no se encuentra agrupado en relación a ninguna variable, debido a la inexistencia de información.

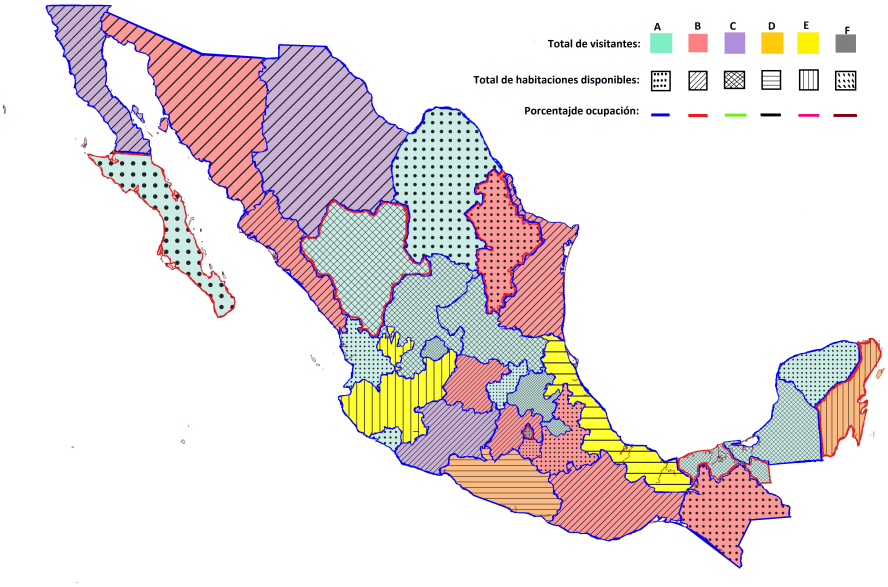


Figura 3.9: Agrupaciones para el sexenio 1, considerando las trayectorias simples.

También se observa que ahora Quintana Roo es el estado que más destaca. Pero Coahuila, Morelos, Tlaxcala y Zacatecas son los estados menos sobresalientes.

Como puede verse en la Figura 3.11 que corresponde al sexenio 3, al igual que en los dos subperiodos anteriores, el número de clústeres coincide al trabajar con las dos variables iniciales, sin embargo para la tercera variable es menor. En esta ocasión el estado que constrasta con en el resto es Quintana Roo y le siguen la Ciudad de México, Guerrero, Jalisco y Veracruz. En este subperiodo, Nayarit, Coahuila, Colima y Yucatán, presentan una mejoría.

En este subperiodo corresponidiente al sexenio 4 (Figura 3.12), Quitana Roo también es el estado que ocupa el nivel más alto en esta actividad (considerando las variables ya mencionadas), los estados subsecuentes son Jalisco, Ciudad de México y Guerrero. Colima, Coahuila y Nayarit ahora se agrupan en clústeres diferentes pero no hay un cambio sobresalientes respecto al periodo anterior a éste. Finalmente, la Figura 3.13 refleja que al considerar el périodo completo, el número de clústeres resultantes es 5, 6, y 6 para cada una de las variables respectivamente, en el orden en el que se han trabajado hasta ahora. También

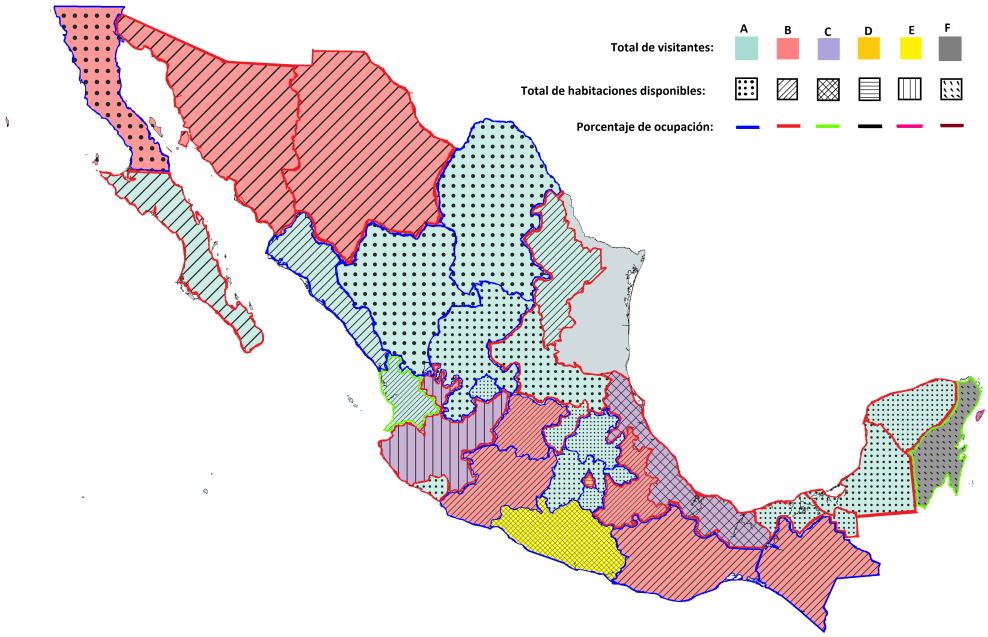


Figura 3.10: Agrupaciones para el sexenio 2, considerando las trayectorias simples.

muestra que como en la mayoría de los subperiodos, Quintana Roo se encuentra posicionado como el estado número uno al considerar las tres variables. Aquí la Ciudad de México se encuentra agrupada en el mismo clúster tomando en cuenta cada una de las variables y también destaca sobre los demás estados.

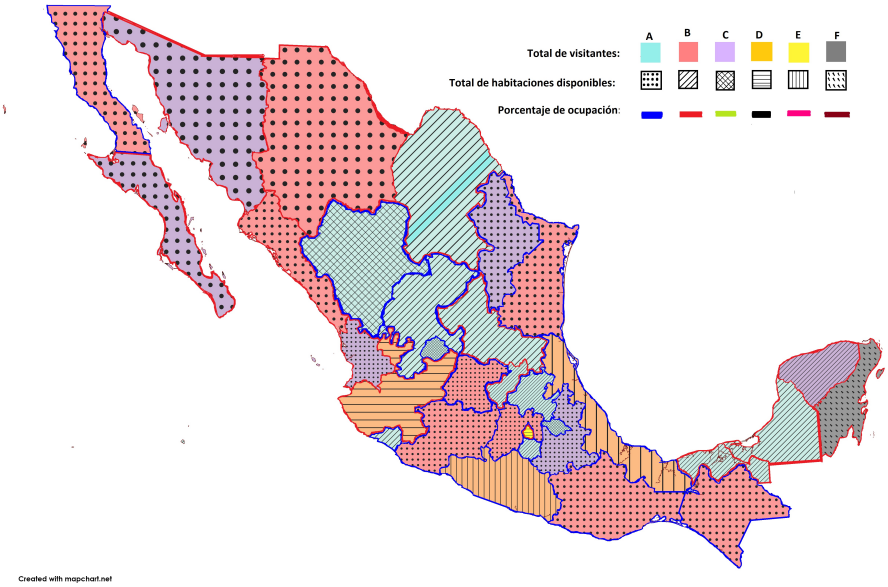


Figura 3.11: Agrupaciones para el sexenio 3, considerando las trayectorias simples.

Análisis de trayectorias conjuntas

Los resultados de las agrupaciones haciendo uso de $kml3d$ se encuentran reportados en la Tabla 3.4 en donde se observan las agrupaciones al considerar periodos sexenales y el lapso comprendido entre enero de 1992 y diciembre de 2019.

Donde S1 denota el subperiodo de 1 de enero de 1992 a 30 de noviembre de 2000, S2 el de 1 de diciembre de 2000 a 30 de noviembre de 2006, S3 va 1 de diciembre de 2006 a 30 de noviembre de 2012, S4 de 1 de diciembre de 2012 a 30 de diciembre de 2019 y T representa el de 1 de diciembre de 1992 a 30 de diciembre de 2019. Y NA en el estado de Tamaulipas y sobre el subperiodo S2 significa que dicho estado no está categorizado, esto debido a la ausencia de información.

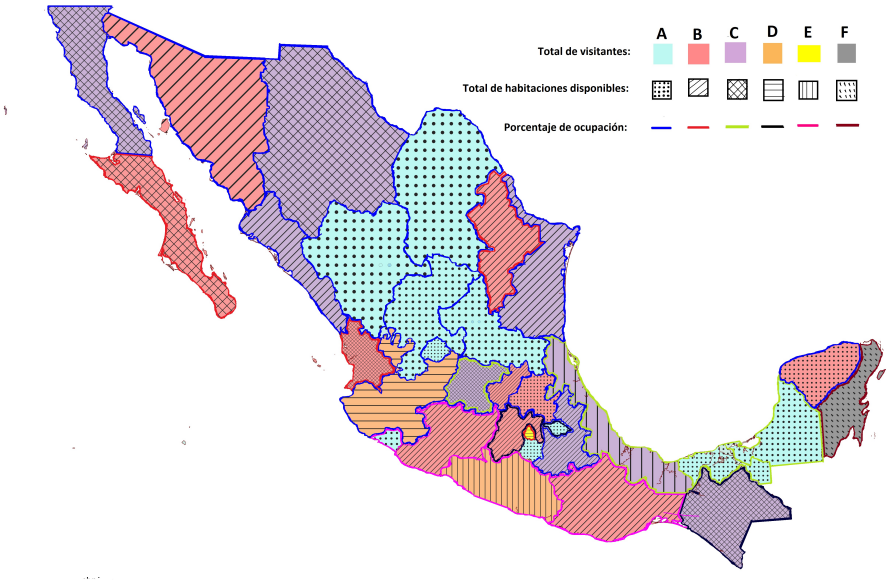


Figura 3.12: Agrupaciones para el sexenio 4, considerando las trayectorias simples.

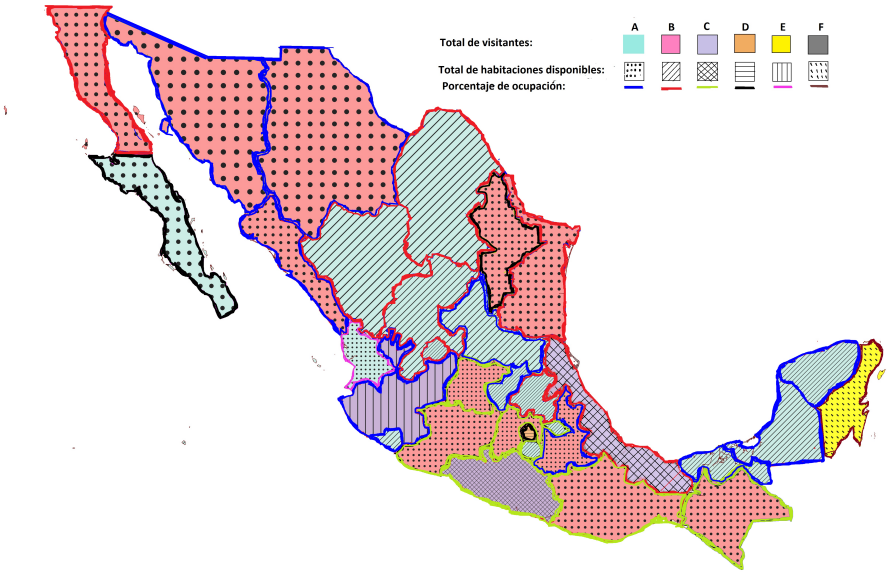


Figura 3.13: Agrupaciones para el periodo completo, considerando las trayectorias simples.

Variables conjuntas					
Estados	S1	S2	S3	S4	T
Aguascalientes	A	A	B	B	B
Baja California Norte	B	B	A	A	A
Baja California Sur	A	B	A	A	A
Campeche	A	A	B	B	B
Chiapas	B	B	A	A	A
Chihuahua	B	B	A	A	A
Ciudad de México	C	C	C	C	D
Coahuila	A	A	B	B	B
Colima	A	A	B	B	B
Durango	A	A	B	B	B
Guanajuato	B	B	A	A	A
Guerrero	C	C	C	A	C
Hidalgo	A	A	B	B	B
Jalisco	C	C	C	C	D
México	B	B	A	A	A
Michoacán	B	B	A	A	A
Morelos	A	A	B	B	B
Nayarit	A	A	A	A	A
Nuevo León	B	B	A	A	A
Oaxaca	B	B	A	A	A
Puebla	B	B	A	A	A
Queretaro	A	A	B	B	B
Quintana Roo	C	C	C	C	D
San Luis Potosí	A	A	B	B	B
Sinaloa	B	B	A	A	A
Sonora	B	B	A	A	A
Tabasco	A	A	B	B	B
Tamaulipas	B	NA	A	A	A
Tlaxcala	A	A	B	B	B
Veracruz	C	C	C	A	C
Yucatán	A	A	B	B	B
Zacatecas	A	A	B	B	B

Tabla 3.4: Distribución de las trayectorias conjuntas según el lapso de tiempo considerado.

Hallazgos del análisis estadístico

Los hallazgos que se desprenden del análisi de clústeres longitudinales se dividirán de acuerdo a la naturaleza de las trayectorias consideradas; simple o conjunta.

Con respecto al análisis de trayectorias simples tenemos que:

- La cantidad de clústeres a lo largo de los cuatro periodos, respecto a la variable "Total de visitantes" y "Total de habitaciones disponibles" coinciden. Mientras que al examinar la variable "Porcentaje de ocupación" la cantidad de clústeres cambia según el periodo de tiempo analizado.

- Aguascalientes, Campeche, Coahuila, Colima, Durango, San Luis Potosí, Tabasco, Tlaxcala y Zacatecas se mantienen en el mismo grupo durante los cuatro sexenios respecto a la variable "Total de visitantes".

- Quintana Roo se encuentra agrupado de forma individual durante los últimos tres periodos. Lo que refleja ser uno de los destinos más apreciados y conocidos de todos los turistas del mundo.

- Ciudad de México, Guerrero, Jalisco y Veracruz también se encuentran entre los estados más visitados.

- Además, Quintana Roo ocupa el primer lugar en disponibilidad de cuartos durante los cuatro periodos estudiados, seguido de la Ciudad de México y Jalisco.

- Quintana Roo también sobresale sobre los demás estados por ser el que cuenta con mayor porcentaje de ocupación a lo largo de los cuatro periodos.

- Al comparar las agrupaciones resultantes en el lapso de 1992 al 2019 y el último sexenio con las agrupaciones resultantes en cada uno de los tres primeros sexenios analizados al considerar la variable "Porcentaje de ocupación", hay un cambio significativo en el número de clústeres. Observe que, este cambio en el número de clústeres es característico únicamente de esta variable.

Con respecto al análisis de trayectorias conjuntas observamos que:

- Al considerar las secciones temporales dadas por los sexenios, se mantiene el agrupamiento en tres clústeres en cada uno de ellos. Sin em bargo, al considerar el periodo de 1992 a 2019, el número de clústeres se eleva a cuatro

- Ciudad de México, Quintana Roo y Jalisco se mantienen agrupados en el mismo clúster a lo largo de los cuatro subperiodos, así como en el periodo bajo estudio (1992-2019). Destacando sobre los demás estados en términos absolutos (sus trayectorias están por encima de las de los demás estados).

- Otros estados que permanecen agrupados en el mismo clúster durante los cinco lapsos (respectivamente) son: agrupados en clústeres iguales (respectivamente) durante los cuatro subperiodos y el periodo completo son:

    - Guerrero y Veracruz.
    - Campeche, Coahuila, Colima, Durango, Hidalgo, Morelos, Querétaro, San Luis Potosí, Tabasco, Tlaxcala, Yucatán y Zacatecas.
    - Chiapas, Chihuahua, Guanajuato, Estado de México, Michoacán, Nuevo León, Oaxaca, Puebla, Sinaloa y Sonora.

- Nayarit se posiciona como el estado que menos sobresale en relación al turismo (en terminos absolutos) situación que se evidencía por la relación que guardan sus trayectorias asociadas con respecto de las del resto de los estados.

Conclusiones

Derivado del análisis longitudinal de trayectorias conjuntas asociadas a las variables "Total de visitantes" , "Total de habitaciones disponibles" y "Porcentaje de ocupación" se observa una clara agrupación de ciertos estados de la República en 4 clústeres. Siendo cada uno de ellos susceptible de estrechar alianzas entre sus miembros y/o conjuntar estrategias promocionales así como políticas públicas comunes en materia de turismo.

Como consecuencia del análisis longitudinal de trayectorias simples asociadas con los estados de la República, se observó una variación en el número de clústeres en la variable "Porcentaje de ocupación" con respecto de las otras dos variables bajo estudio. Lo cuál puede interpretarse como evidencia de las fluctuaciones del desplazamiento de turistas a diversos puntos del país como resultado (posiblemente) de promociones y/o políticas locales o modas pasajeras.

Con respecto a la metodología de análisis de clústeres longitudinales se concluye que es una herramienta potente y versátil para clasificar entidades basados en trayectorias de series de tiempo y que por tanto es un apoyo invaluable para detectar similitudes entre ellas, lo que puede conllevar a la toma de decisiones, que además cuenta con el complemento computacional ofrecido por el software libre, situación que vuelve al análisis de clústeres longitudinales en un conjunto de técnicas altamente aplicables por parte de investigadores de distintas disciplinas.

Con respecto al trabajo futuro, se abre una oportunidad para replicar el estudio presentado en este trabajo de Tesis en fenómenos cuyas bases de datos asociadas se encuentren organizadas como trayectorias de series de tiempo, así como el estrechamiento de cooperaciones con investigadores de diversas disciplinas para analizar estadísticamente la agrupación de entes a través del tiempo cuando así sea requerido.

Apéndice A

Códigos en R

A continuación se presentan algunos códigos ejecutados en **R** para los análisis repotados en el capítulo 3. Se presenta sólo un caso de análisis de trayectorias simples y uno para trayectorias conjuntas, ya que los demás análisis se hacen de manera análoga.

Para hacer el análisis de la variable "Total de visitantes":

```
MV1 <- read_excel("C:/Users/ARE/Downloads/RparaTesis/MV1.xlsx")
View(MV1)

knitr::kable(MV1 %>% head(5))
unique(MV1$Estado) %>%  length()
is.na(MV1) %>% sum()

MV1d<- as.data.frame(MV1 )

set.seed(1234)
MV1_cld<-kml::cld(MV1d, timeInData = 2:337)
MV1_cld

(option1 <- parALGO())
result<-kml(MV1_cld,nbRedrawing=20,toPlot="traj", parAlgo=option1)
```

```
X11(type="Xlib")
plotAllCriterion(MV1_cld)

X11(type="Xlib")
try(choice(MV1_cld))

result$clust <- getClusters(MV1_cld,5)
result$clust
```

A continuación se presenta el código que se usó para el análisis de las tres variebles en conjuntas en el sexenio 1:

```
Lap1Vc <- read_excel("C:/Users/HP/Downloads/RparaTesis/Lap1Vc.xlsx")
View(Lap1Vc)
Lap1Vcd<- as.data.frame(Lap1Vc)
head(Lap1Vcd)

set.seed(1234)
cldLap1Vcd<- kml3d::clusterLongData3d(Lap1Vcd, timeInData = list(v1= 2:
108, v2 = 109:215, v3= 216:322 ))
cldLap1Vcd
(option1 <-parKml3d())
result<-kml3d(cldLap1Vcd,nbRedrawing=20,toPlot="both", parAlgo=option1)

X11(type="Xlib")
try(plotAllCriterion(cldLap1Vcd))

X11(type="Xlib")
try(choice(cldLap1Vcd))

result$clust <- getClusters(cldLap1Vcd ,2)
result$clust
```

Bibliografía

[1] B. S. Everitt, S. Landau, M. Leese and D. Stahl, *Cluster Analysis, 5th Edition*. John Wiley y Sons, 2011.

[2] C. Genolini and B. Falissard. *Kml:K-means for longitudinal data*. Computational Statistics, 25(2):317-328, 2010.

[3] C. Genolini, X. Alacoque, M. Sentenac and C. Arnaud : *kml and kml3d: R Packages to Cluster Longitudinal Data.* Journal of Statistical Software, 65(4), 1-34. doi:10.18637/jss.v065.i04 <https://doi.org/10.18637/jss.v065.i04>, <https://www.jstatsoft.org/article/view/v065i04>.

[4] D. Peña. *Análisis de datos Multivariantes*. McGraw-Hill Interamericana de España S.L., 2002.

[5] D. Phung, G. I. Webb and C. Sammut. *Encyclopedia of machine learning and data science*. Springer, 2020.

[6] D. Teuling and V. den Heuvel. *Clustering of longitudinaldata: a tutorial on a variety of approaches* .Springer, 2020.

[7] E. Crivisqui. *Notas de curso*. Programa PRESTA, 1999.

[8] E. Propín and A. Sánchez. *Tipología de los municipios turísticos de México a fines del siglo XX*. Geographicalia,1998.

[9] F. Husson, S. Le and J. Pagès. *Exploratory Multivariate Analysis by Example Using R*. Chapman and Hall/CRC; 2nd edición , 2020.

[10] F. Sánchez. *Turismo receptivo y crecimiento económico en México: evidencia a largo plazo*. Accounting y Management, 2020.

[11] G. Bacallao. (2023, Mayo 10) *Encyclopedia Britannica*[Versión electrónica]. Available: *https://www.britannica.com/topic/cluster-analysis.*

[12] H. Wickham, J. Hester, W. Chang and J. Bryan. _*devtools*(2022)*: Tools to Make Developing R Packages Easier_. R package version 2.4.5, <https://CRAN.R-project.org/package=devtools>.*

[13] INEGI: *Cuenta Satélite del Turismo de México: https:/ww.inegi.org.mx/app/saladeprensa/noticia.html?id=7874.*

[14] J. K. Mandal, D. Bhattachaya. *Emergin Technology in modelling and graphics*. Springer, 2018.

[15] J. R. Demey, L. Pla, J. L. Vicente-Villardón, J. A. Di Rienzo and F. Casanoves. *"Medidas de distancia y de similitud" en Valoración y análisis de la diversidad funcional y su relación con los servicios ecosistémicos*. Costa Rica: CATIE, 2011.

[16] L. Kaufman and P. J. Rousseeuw. *Finding groups in data, an introduction to cluster analysis*. Wiley-Interscience, 1990.

[17] L. Nalvarte. *Análisis multivariado I, clúster jerárquico*. Facultad de Ciencias Económicas y de Administración, Universidad de la República , 2021.

[18] M. Chavent, Y. Lechevallier and O. Briant. *DIVCLUS-T: a monothetic divisive hierarchical clustering method*. Computational Statistics an Data Analysis, 2007, 52(2), pp.687-701.10.1016/j.csda.2007.03.013. hal-00260963.

[19] M. Gutiérrez- Lagunes and N. H. Monroy. *Un estudio exploratorio del sector turismo en México*. Congreso internacional de contaduría, adminsitración e informática, septiembre 2018.

[20] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert and K. Hornik, . *cluster (2022): Cluster Analysis Basics and Extensions. R package version 2.1.4.*

[21] M. O. Lobo, C. A. Flores, J. Quiroz and I. Cruz. *Factors that affect the demand of tourism in Mexico: competitive analysis.* Journal of Tourism Analysis, Vol. 25 No. 2, pp. 154-166, 2018.

[22] M. Sánchez. *Análisis cuantitativo del impacto económico de la competitividad en destinos turísticos internacionales.* Revista de economía Mundial, 2012.

[23] Organización Mundial del Turismo. *Glosario de terminos del turismo: https://www.unwto.org/es/glosario-terminos-turisticos.*

[24] P. Diggle, P. Heagerty, K.Y Liang and S. Zeger. *Analysis on longitudinal Data(second edition)*, Oxford: Oxford University Press, 2002.

[25] P. Giordani, M.B. Ferraro and F. Martella. *An introduction to clustering with R*, Springer, 2020.

[26] R Core Team (2022). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.*