UNA PROPUESTA DE REGIONALIZACIÓN DE LA REPÚBLICA MEXICANA CON BASE EN INDICADORES ELECTORALES.

Víctor Florentino Miranda Soberanis

A Dios, por regalarme la vida;

y a Shender, por su infinito amor.

Directora de Tesis: Dra. Silvia Ruiz-Velasco Acosta

Enero del 2008

0.1. Agradecimientos

No ha sido fácil y/o sencillo llegar hasta este punto de mis metas profesionales, pero ¿quién estipuló que así sería? Es una pregunta cuya respuesta he ido madurando con el paso del tiempo. Este camino inició en el mes de agosto del año 2005, con una actitud a la *expectativa* y con cierto nerviosismo, pero con la firme decisión de culminar con este proyecto de titulación de maestría.

Primero y por sobre todas las cosas quiero agradecer a Dios, por haberme regalado la salud y por haber colocado en mi camino a todas aquellas personas quienes, directa o indirectamente, contribuyeron para terminar este proyecto.

A tí mi amada Shender por tu eterno amor, tu infinita paciencia, tu perenne comprensión y tu incansable apoyo. Porque con el solo hecho de estar en mi corazón, haces que todo valga la pena.

A mi madre y familia materna por su apoyo e impulso, que fueron determinantes para poder continuar: Rita del Carmen, Helga, Pilar, Adda Beatriz, Víctor Hugo, Felipe, Francisco, Antonio, tíos políticos y primos. Asimismo, hubo gente amable, que en cada momento ofrecieron un soporte incondicional: las familias Ávila Sansores y Ávila Chan.

Sin embargo, no es necesario estar físicamente presente para sentir un verdadero apoyo. Dedico este trabajo a mis abuelos Pilar y Víctor, y mi tío Ignacio, pues desde el cielo estoy seguro que me enviaron bendiciones y quienes, en vida, me aconsejaron para optar por el camino correcto.

También quiero agradecer a la Universidad Autónoma de Yucatán por mi formación profesional, y de manera muy especial a dos profesoras por su comprensión, consejo y apoyo invaluable para seguir adelante en momentos difíciles de mi vida: la M.C. Lucy Torres Sanchez y la I.Q.I Irene Peniche Ayora.

Ahora, refiriéndome a las personas que contribuyeron de manera directa para la realización de este proyecto. En primer lugar a la Universidad Nacional Autónoma de México (UNAM) y el CONACyT por ser las instituciones que me brindaron la oportunidad de realizar mis estudios de maestría y por contar con una excelente plantilla de profesores. Agradezco a la Dra. Silvia Ruiz -Velasco Acosta por su paciencia, dedicación y enseñanzas, quien fue mi directora de tesis; al Dr. Federico O′Reilly Togno, quien fue mi tutor de la maestría y me orientó en todo momento, siempre recordaré su curso de Modelos Lineales; y al Dr. Ignacio Méndez Ramírez por sus excelentes consejos y orientación en la maestría y futuros proyectos profesionales.

Agradezco también a mis sinodales, quienes con paciencia revisaron este proyecto: La M.C. Leticia Gracia-Medrano Valdelamar por su conocimiento y recomendaciones para mejorarlo, la M.C. Patricia Romero Mares por su claridad en la redacción y demostraciones, y el Dr. Ramsés Mena Chávez por su objetividad en aspectos matemáticos y recomendaciones en cuanto al procesador de textos.

Finalmente quiero mencionar que en este tiempo no solo me formé profesionalmente y aprendí con los estudios de licenciatura y maestría, pues hubo momentos difíciles que hicieron que cada vez valorara más esta oportunidad de estudiar, que borrara los defectos de soberbia y prepotencia, y pusiera en práctica los principios de humildad, disciplina, orden, responsabilidad y Fé. Aprendí algo mucho más importante: **El intelecto y la humildad pueden ser compatibles, toda vez que la humildad esté en primer lugar**.

Índice general

0.2. Introducción

A partir de 1994, el Registro Federal de Electores ha llevado a cabo diversos ejercicios cuyo objetivo es evaluar la calidad del Padrón Electoral con base en métodos estadísticos. Estos trabajos incluyen:

- Verificación Nacional de 1994.
- Diagnóstico muestral del Padrón Electoral de 1996.
- Verificación Nacional de 1997.
- Verificación Nacional Muestral al Padrón Electoral del año 2000.
- Verificación Nacional Muestral 2002.
- Verificación Nacional Muestral 2003.
- Verificación Nacional Muestral 2005.
- Verificación Nacional Muestral 2006

En particular, el propósito de estos ejercicios electorales ha sido proporcionar al Consejo General del Instituto Federal Electoral elementos de juicio para que emita su veredicto sobre la validez del Padrón Electoral y la Lista Nominal que son empleados en las elecciones federales. El diseño estadístico ha permitido hacer inferencias a nivel nacional, estatal y regional[1], y los resultados contribuyen al mejoramiento de programas de empadronamiento y credencialización, así como de depuración del Padrón Electoral. En estas evaluaciones se ha incluido la realización de dos encuestas: Cobertura y Actualización; y han sido supervisadas por los miembros de la Comisión Nacional de Vigilancia.

[1] El nivel de estas inferencias no siempre ha sido el mismo, es decir, ha variado en función de aspectos como el tiempo, presupuesto, etc.

La encuesta de Cobertura tiene la intención de calificar el empadronamiento de la población residente que tiene 18 años de edad o más. Para tal efecto se evalúa el hecho de que las personas estén empadronadas, que posean su Credencial para Votar, que su empadronamiento y credencial tenga sus datos correctos, y que correspondan a la sección electoral y/o domicilio en donde residen. La encuesta de Actualización tiene como objetivo general constatar que cada registro del Padrón Electoral y de la Lista Nominal corresponda a un ciudadano. Estas evaluaciones se pueden clasificar en dos grupos: **a)** las que tienen como objetivo primordial la evaluación, y **b)** las que tienen como objetivo principal proveer datos para el mejoramiento del padrón o del empadronamiento.

Entre las primeras están las Verificaciones de 1994, 1997 y 2000. Éstas han sido realizadas para evaluar el Padrón Electoral previo a las elecciones federales, y fueron realizadas pocos meses antes de la fecha electoral. Con los indicadores que se han obtenido de ellas, el Consejo General del IFE ha tenido un apoyo relevante para declarar al Padrón como un instrumento válido y definitivo para la realización de las elecciones de esos años. Al mismo tiempo, la Comisión Nacional de Vigilancia tiene elementos de juicio para procurar que el Padrón y Lista Nominal se hayan elaborado con transparencia, suficiencia, y bajo los lineamientos que establece el Código Federal de Instituciones y Procedimientos Electorales.

Entre las segundas están las de 1996 y de 2002. Éstas se han realizado el año previo a la fecha electoral con la finalidad de conocer la situación del empadronamiento, previo a las elecciones. Su objetivo ha sido evaluar y proporcionar datos para mejorar la calidad del empadronamiento y del Padrón Electoral.

Por otro lado, los diferentes esquemas de muestreo hicieron que el alcance de la inferencia (nacional, regional, etc.) no siempre fuera el mismo. De hecho, en varios de estos trabajos el muestreo se realizó en dos estratos: secciones urbanas y secciones no urbanas (1997, 2000, 2003, 2005); y más aún, fue en la Verificación Nacional Muestral del 2000 cuando por primera vez se realizaron inferencias a nivel regional. Para tal efecto, se revisaron varios trabajos de regionalización y por las características de las entidades en cuestión, se consideró que la propuesta más conveniente sería la previamente mencionada regionalización elaborada por CONAPO, basada en el Índice de Marginación.

El Índice de Marginación (IM) es una medida-resumen que permite diferenciar entidades federativas y municipios según el impacto global de las carencias que padece la población, como resultado de la falta de acceso a la educación, la residencia en viviendas inadecuadas, la percepción de ingresos monetarios insuficientes para adquirir una canasta básica y las relacionadas con la residencia en localidades pequeñas.

El objetivo general de este proyecto es proponer una regionalización de los Estados Unidos Mexicanos a nivel estatal, que sea útil principalmente para efectos electorales, así como la realización de inferencias a nivel regional o bien, la estimación de indicadores con mayor precisión. Para tal efecto, se utilizarán los datos que arrojaron las Verificaciones Nacionales Muestrales de 2005 y 2006, en particular, información sobre los indicadores electorales considerados en éstas. Se hará uso de estas fuentes por las siguientes ventajas:

- la información sobre los indicadores es reciente,

- mantienen el marco conceptual del estudio y
- se tiene información a nivel entidad federativa por cada indicador.

Se utilizarán herramientas de Estadística Multivariada que permitan discriminar la información en grupos homogéneos y que al mismo tiempo sean exhaustivas. En este sentido, las regiones (conformadas por entidades federativas) no necesariamente serán conexas, pero serán homogéneas en cuanto a características electorales reflejadas en los principales indicadores. La motivación para la realización de este proyecto es que la regionalización resultante será más coherente para estudios con fines electorales (estadísticos, socio - demográficos, etc.) y en los cuales se requieran inferencias o conclusiones a nivel regional.

En el primer capítulo se describirán las Verificaciones Nacionales Muestrales que hasta ahora se han realizado y que juegan el papel de antecedentes, con la intención de hacer notar la motivación de este proyecto. En el segundo capítulo se presenta una construcción detallada del Índice de Marginación de 1990 calculado por CONAPO y que dió lugar a la regionalización utilizada en la verificación del año 2000, con la intención de hacer énfasis en que dicha regionalización no fue realizada con fines electorales y por ende, las inferencias[2] (con fines electorales) para dichas regiones pueden no reflejar la realidad.

En el tercer capítulo se describirán detalladamente las principales técnicas multivariadas a utilizar; para aplicarlas se hará uso del software STATA versión 9.2 y S-PLUS V6.1. En el cuarto capítulo se exhibirán los resultados obtenidos con los datos de las Verificaciones de 2005 y 2006. En el capítulo cinco se realizará el análisis y validación de las regiones propuestas, principalmente con un

[2]Nacional, estatales y/o regionales.

Análisis Discriminante. Asimismo, se exhibirán medidas resumen de dichas regiones propuestas con el fin de describirlas en sus principales indicadores electorales. Finalmente se presentarán tres anexos; en los dos primeros se detallarán las diferentes técnicas alternas que se mencionarán y/o emplearán en este proyecto. En el tercer anexo se definirán los indicadores electorales cuyos valores serán objeto de análisis.

Para un buen entendimiento, se recomienda al lector tener conocimientos sólidos de álgebra lineal, estadística multivariada y cálculo en varias dimensiones; así como conceptos básicos sobre la creación de índices socio-demográficos.

Capítulo 1

Antecedentes y Motivación del Proyecto

1.1. El Instituto Federal Electoral.

El artículo 41, Fracción III de la Constitución Política de los Estados Unidos Mexicanos, establece que la organización de las elecciones federales es una función estatal que se ejerce a través del Instituto Federal Electoral (IFE) como un organismo público autónomo, dotado de personalidad jurídica y patrimonio propios.

Bajo este contexto y desde su creación, el IFE ha estado comprometido en garantizar que la organización de los procesos electorales se desarrolle con estricto apego a sus principios rectores: certeza, legalidad, independencia, imparcialidad y objetividad, que comprenden una concepción del deber ser del Instituto y de quienes lo integran, con el fin de dar cumplimiento a la función que le ha sido encomendada.

En este sentido, el artículo 69 del Código Federal de Instituciones y Procedimientos Electorales (COFIPE) apunta como fines del Instituto Federal Electoral:

- Contribuir al desarrollo de la vida democrática,

- Preservar el fortalecimiento del régimen de partidos políticos,
- Integrar el Registro Federal de Electores,
- Asegurar a los ciudadanos el ejercicio de los derechos político-electorales y vigilar el cumplimiento de sus obligaciones,
- Garantizar la celebración periódica y pacífica de las elecciones para renovar a los integrantes de los Poderes Legislativo y Ejecutivo de la Unión,
- Velar por la autenticidad y efectividad del sufragio,
- Llevar a cabo la promoción del voto y coadyuvar a la difusión de la cultura democrática.

Los instrumentos electorales que el IFE proporciona a la sociedad mexicana, ciudadanos y actores políticos, constituyen un factor esencial para garantizar procesos electorales legítimos, representativos y confiables; además, dan sustento al Sistema Electoral Mexicano en cuanto a los principios de legalidad, representatividad y certeza. Estos instrumentos son: el Padrón Electoral, la Credencial para Votar, la Lista Nominal de Electores y la Cartografía Electoral.

- El **Padrón Electoral** es el listado en el que se encuentran todos los ciudadanos mexicanos que solicitaron su inscripción al mismo, con la finalidad de obtener su Credencial para Votar con fotografía y así poder ejercer su derecho al voto. El Padrón Electoral actual surgió a partir de la aplicación de la técnica censal total en 1991 como parte del programa "Depuración Integral del Padrón Electoral y Nueva Credencial para Votar con Fotografía", con miras a ser utilizado en el proceso electoral de 1994. Actualmente, el Padrón Electoral se encuentra dividido en 63,634 secciones a nivel nacional y es un instrumento que goza de la credibilidad y confianza de la ciudadanía

y de los actores políticos, gracias a los procesos de depuración y actualización que el Registro Federal de Electores ha llevado a cabo.

- La **Lista Nominal** es el listado que se utiliza en la casilla el día de la jornada electoral que incluye la fotografía y el nombre de los ciudadanos registrados en el Padrón Electoral y que ya obtuvieron su Credencial para Votar con Fotografía. La finalidad en su utilización es cotejar los datos del ciudadano en el ejercicio del voto.

- La **Cartografía Electoral** es una técnica informativa que tiene por objeto levantar, redactar y publicar mapas del estado, municipios, distritos, secciones y manzanas para una inmediata localización. A través de la cartografía electoral se realiza la representación gráfica del marco geográfico-electoral, además de dar a conocer la distribución de los ciudadanos con derecho al sufragio en el territorio nacional.

  Mantener actualizada la Cartografía Electoral es una tarea permanente y complicada debido a la complejidad del territorio nacional en zonas urbanas y en las localidades en zonas rurales y mixtas; además de los movimientos demográficos y la geografía física propia de las circunscripciones de estudio. Cabe mencionar que hasta el 2003, la mayor concentración del Padrón Electoral se encontraba en zonas urbanas (69 %), y se concentraba principalmente en los siguientes estados: Distrito Federal, Estado de México, Jalisco, Nuevo León y Veracruz.

1.2. La Dirección Ejecutiva del Registro Federal de Electores.

El Instituto Federal Electoral por su parte, proporciona estos instrumentos a través de la Dirección Ejecutiva del Registro Federal de Electores (DERFE), el cual es un órgano que se reconoce como parte del mismo Instituto y cuya función social radica en proveer los instrumentos electorales a los ciudadanos mexicanos y a los actores políticos de manera segura, oportuna y confiable; buscando contribuir a la transparencia y credibilidad de los procesos electorales.

El Código Federal de Instituciones y Procedimientos Electorales establece en su artículo 92, las atribuciones conferidas al Registro Federal de Electores (mismas que se enlistan a continuación), y que fundamentan el diseño y realización de las actividades y responsabilidades de la Dirección Ejecutiva del Registro Federal de Electores.

- Formar el Catálogo General de Electores,
- Aplicar, en los términos del artículo 141 de este Código, la técnica censal total en el territorio del país para formar el Catálogo General de Electores,
- Aplicar la técnica censal en forma parcial en el ámbito territorial que determine la Junta General Ejecutiva,
- Formar el Padrón Electoral,
- Expedir la Credencial para Votar según lo dispuesto en el Título Primero del Libro Cuarto de este Código,
- Revisar y actualizar anualmente el Padrón Electoral conforme al procedimiento establecido en el Capítulo Tercero del Título Primero del Libro

Cuarto de este Código,

- Establecer con las autoridades federales, estatales y municipales la coordinación necesaria, a fin de obtener la información sobre fallecimientos de los ciudadanos, o sobre pérdida, suspensión u obtención de la ciudadanía,
- Proporcionar a los órganos competentes del Instituto y a los partidos políticos nacionales, las listas de electores en los términos de este Código,
- Formular, con base en los estudios que realice, el proyecto de división del territorio nacional en 300 distritos electorales uninominales, así como el de las cinco circunscripciones plurinominales,
- Mantener actualizada la Cartografía Electoral del país, clasificada por entidad, distrito electoral federal, municipio y sección electoral,
- Asegurar que las comisiones de vigilancia nacional, estatales y distritales se integren, sesionen y funcionen en los términos previstos por este Código,
- Llevar los libros de registro y asistencia de los representantes de los partidos políticos a las comisiones de vigilancia,
- Solicitar a las comisiones de vigilancia los estudios y el desahogo de las consultas sobre los asuntos que estime conveniente dentro de la esfera de su competencia,
- Las demás que le confiera este Código.

Bajo esta óptica, el Código Federal de Instituciones y Procedimientos Electorales:

1. Define la estructura que mejor permita al Instituto Federal Electoral cumplir con el mandato constitucional y que se conforma por órganos de dirección, ejecutivos técnicos y de vigilancia.

2. Señala que las actividades del Instituto Federal Electoral deberán estar encaminadas hacia el logro de determinados fines institucionales (previamente señalados), los cuales conforman su razón de ser, su mandato, su compromiso y el papel que debe cumplir para con la sociedad mexicana.

1.3. La Comisión Nacional de Vigilancia.

En referencia al punto número uno, se creó la Comisión Nacional de Vigilancia que es un órgano integrado mayoritariamente por representantes de los partidos políticos, que coadyuva en los trabajos relativos a la actualización del Padrón Electoral y realiza funciones de vigilancia en las actividades de la DERFE contenidas en el Libro Cuarto del COFIPE. Tiene su *fundamento legal* en la Fracción III, párrafo segundo del artículo 41 de la Constitución Política de los Estados Unidos Mexicanos, párrafo 2 del artículo 92 del Código Federal de Instituciones y Procedimientos Electorales y artículos 165 y 166 del mismo ordenamiento.

La Comisión Nacional de Vigilancia se instaló el 21 de noviembre de 1990, con la asistencia de los partidos PAN, PRI, PPS, PRD, PFCRN, PARM y PDM. Se integra por un Presidente que es el titular de la Dirección Ejecutiva del Registro Federal de Electores, un representante propietario y su suplente por cada uno de los partidos políticos nacionales, y un Secretario, el cual es designado por el Presidente del personal del Servicio Profesional Electoral. Asímismo cuenta con la participación de un representante del Instituto Nacional de Estadística, Geografía e Informática.

Este órgano máximo de vigilancia ha abordado, entre otros, los siguientes temas:

- Modificaciones de plazos para credencialización,
- Programas de depuración del Padrón Electoral,
- Programas de fotocredencialización y diseño del modelo de la credencial para votar con fotografía, de listas nominales para exhibición y definitivas,
- Metodologías y aplicación de verificaciones y diagnóstico al Padrón Electoral,
- Planeación de las campañas de actualización al padrón electoral intensas y permanentes: aplicación del artículo 163 del COFIPE en entidades con proceso electoral local,
- Asuntos relativos a la campaña de difusión de los programas de la Dirección Ejecutiva del Registro Federal de Electores, así como la atención de la ciudadanía a través de los Centros Estatales de Consulta Electoral y Orientación Ciudadana e IFETEL.

1.4. La Comisión Nacional de Supervisión y Evaluación.

Paralelamente, los representantes de los partidos políticos consideraron pertinente crear un órgano técnico que los apoyara tanto en sus actividades como en las del Comité Nacional de Vigilancia, éste fue denominado Comité Nacional de Supervisión y Evaluación que es el órgano técnico auxiliar de la Comisión Nacional de Vigilancia en el cual se originan y formulan las propuestas de planeación, programación, supervisión, evaluación, seguimiento y auditoría de los trabajos de actualización y depuración del Padrón Electoral. Tiene su *fundamento legal* en el Acuerdo 2-34 : 11/12/92 de fecha 11 de diciembre de 1992 y acuerdo 2-57 : 24/11/94 del 24 de noviembre de 1994, de la Comisión Nacional de Vigilancia.

El Comité Nacional de Supervisión y Evaluación está integrado por un Coordinador General Técnico del Registro Federal de Electores, quien se auxilia de los funcionarios de la DERFE que requiera para el desarrollo de esta función; el Secretario es el titular de la Secretaría de Comisión Nacional de Vigilancia; ambos asisten únicamente con derecho a voz.

Adicionalmente, se cuenta con dos representantes por cada partido político: uno es el representante ante la Comisión Nacional de Vigilancia y el otro es designado por éste; ambos con sus respectivos suplentes, y deben contar con una alta calificación técnica, fundamentalmente en las áreas de estadística, actuaría, informática o demografía. A la Comisión Nacional de Supervisión y Evaluación le compete analizar, principalmente, los siguientes asuntos:

1. La actualización de la Cartografía Electoral,
2. Planeación a detalle de las campañas de actualización,
3. Diseño de las campañas de difusión,
4. Seguimiento integral de las campañas de actualización del Padrón Electoral,
5. Depuración y mejoramiento de la calidad del Padrón Electoral,
6. Diagnóstico al Padrón Electoral y verificación nacional muestral del Padrón Electoral.

Dentro de la estructura del Instituto Federal Electoral se encuentra la Secretaría de la Comisión Nacional de Vigilancia cuyo objetivo es coordinar las actividades de la Comisión Nacional de Vigilancia y del Comité Nacional de Supervisión y Evaluación, así como orientar y apoyar a las Comisiones Locales y

Distritales de Vigilancia en el desarrollo de sus actividades, vigilar el cumplimiento de sus acuerdos e integrar la información generada por los órganos de vigilancia y las áreas de la Dirección Ejecutiva del Registro Federal de Electores. Entre las funciones de dicha Secretaría están:

- Supervisar que los órganos de vigilancia se integren, sesionen y funcionen en los términos previstos por el Código Federal de Instituciones y Procedimientos Electorales.
- Coordinar las sesiones y los trabajos de la Comisión Nacional de Vigilancia y del Comité Nacional de Supervisión y Evaluación, así como de los grupos de trabajo que se formen para analizar y dar cumplimiento a los compromisos establecidos.
- Concertar con los representantes de los Partidos Políticos, las actividades de la Comisión Nacional de Vigilancia y del Comité Nacional de Supervisión y Evaluación.
- Coordinar la administración y entrega de los recursos económicos asignados a la Comisión Nacional de Vigilancia y el Comité Nacional de Supervisión y Evaluación.
- Coordinar la elaboración de los informes a nivel nacional, para su análisis y registro en cuanto a asistencia a las reuniones, acuerdos adoptados y acreditación de los representantes de los partidos políticos.

1.5. Verificaciones Muestrales a nivel nacional.

Dentro de los trabajos de actualización y diagnóstico al Padrón Electoral y dando continuidad a los ejercicios muestrales que tienen como propósito evaluar su calidad, de 1994 a 2006 la Dirección Ejecutiva del Registro Federal de Electores, bajo la supervisión de la Comisión Nacional de Vigilancia, ha realizado ocho auditorías al Padrón Electoral basadas en métodos de muestreo probabilístico. El objetivo general ha sido conocer tanto el grado de certeza y confiabilidad de los datos incluidos en el Padrón Electoral y la Lista Nominal de Electores así como las condiciones de empadronamiento de los ciudadanos residentes en el país.

El diseño conceptual y estadístico de dichas auditorías ha variado conforme las necesidades de información; sin embargo, se ha ido consolidando un esquema de *revisión muestral* que consiste en la realización de dos encuestas: la Encuesta de Cobertura y la Encuesta de Actualización. La primera tiene como objetivo conocer la situación de empadronamiento de los ciudadanos residentes en el país, mientras que la segunda busca conocer la situación de los registros de la base de datos del padrón. Estas revisiones, denominadas "Verificaciones Muestrales", se dividen en dos tipos en función de la utilidad inmediata de los resultados:

1. Aquellas realizadas en años electorales, con la finalidad de proporcionar a los actores políticos indicadores sobre la calidad de los instrumentos electorales que serán empleados en las elecciones federales. En este grupo quedan clasificadas las evaluaciones realizadas en los años de 1994, 1997, 2000, 2003 y 2006.

2. Aquellas que se realizan en años preelectorales, y cuyo objetivo es apoyar la planeación de campañas de empadronamiento, credencialización y depu-

ración del padrón que anteceden a las elecciones federales. En este grupo quedan clasificadas las evaluaciones realizadas en 1996, 2002 y 2005.

Verificaciones Muestrales realizadas en años electorales

1.5.1. Verificación Nacional Muestral 1994.

La Comisión Nacional de Vigilancia en su sesión del 25 de noviembre de 1993 acordó realizar la Verificación Nacional de 1994, la cual incluyó entre sus objetivos el siguiente: Comprobar la existencia del ciudadano que realizó algún trámite ante la Dirección Ejecutiva del Registro Federal de Electores, ya sea para inscribirse por primera vez, actualizar sus datos de domicilio, o bien corregir información personal. El marco muestral fue el padrón que se tenía en ese momento, el método de selección fue el muestreo aleatorio simple con selección sistemática de registros en los padrones estatales. El nivel de confianza se fijó en 98 % y la precisión a nivel nacional fue 0.6 %. Los resultados se presentaron en dos vertientes[1]:

1. Verificación documental del registro del ciudadano y
2. Verificación en campo.

Para el primer punto:

Muestra nacional = 82,405 registros
Cobertura: Documental
Registros por verificar: 82,256 (100 %)
Registros verificados: 82,213 (99.9 %)

[1] IFE, RFE. "Verificación Nacional Muestral al Padrón Electoral del año 2000. Informe final". 23 de mayo del 2000.

Para el segundo punto:

Cobertura: Campo
Registros por verificar: 75,314 (100 %)
Registros Verificados: 75,054 (99.9 %)

1.5.2. Verificación Nacional Muestral 1997.

En 1997 la Comisión Nacional de Vigilancia solicitó un estudio de verificación muestral del Padrón Electoral y lista nominal que determinara el nivel de cobertura, actualización y consistencia de dichos instrumentos electorales. El marco muestral se formó con las secciones electorales, el tipo de muestreo adoptado fue polietápico, donde las unidades primarias fueron las secciones y las secundarias las manzanas o localidades. El nivel de confianza se fijó en 95 % y la precisión en 2 %. Los resultados, al igual que en la Verificación Nacional Muestral 1994, se presentaron en dos vertientes [2]:

Cobertura

Cobertura del padrón: 90.8 %
Vigencia del padrón: 71.8 %
Cobertura de la credencial: 86.4 %
Vigencia de la credencial: 69.0 %

Actualización

Desactualización del padrón: 20.5 %
Rezago del padrón: 28.2 %

[2] IFE, RFE. "Verificación Nacional Muestral al Padrón Electoral del año 2000. Informe final". 23 de mayo del 2000.

1.5.3. Verificación Nacional Muestral 2000.

El tema de la VNM2000 se abordó en la sesión extraordinaria del Comité Nacional de Supervisión y Evaluación (CONASE), celebrada los días 13 y 14 de septiembre de 1999 en la ciudad de Toluca adoptándose el acuerdo E-002-99 para la ejecución de dicho ejercicio. La participación de los partidos políticos en esta verificación fue destacada ya que intervinieron en todos los aspectos del diseño de trabajo, a decir,

1. la validación y la aprobación de los elementos básicos para el diseño y selección de la muestra,
2. el establecimiento de la normatividad para el trabajo de campo,
3. la conformación de los cuadros operativos.

Ahora bien, desde el punto de vista estadístico y conceptual, la estrategia seguida en la VNM2000 fue diseñar dos encuestas levantadas mediante visitas domiciliarias: una orientada a verificar la calidad de los registros de la base de datos del padrón llamada *encuesta de actualización*; y la segunda dirigida a obtener información sobre la situación del registro electoral de la población de 18 años o más, denominada *encuesta de cobertura y vigencia*[3].

El desarrollo de los Estudios de Caso, que incluye la realización de dos encuestas, complementó la revisión del padrón electoral durante la VNM2000. El objetivo de dichas encuestas fué conocer la institución y/o persona que recogió la credencial para votar y con qué fines, así como la fecha en que ésta fue recogida al ciudadano y la fecha de la devolución.

[3] Adosado a ésta se habilitó un cuestionario para obtener datos sobre los ciudadanos que declaraban no tener credencial de elector debido a que se les habían recogido.

1.5.3.1. Objetivos.

Para la VNM2000 se fijó el siguiente como objetivo general: Establecer estadísticamente el nivel de consistencia del Padrón Electoral, manteniendo la comparabilidad conceptual con la verificación de 1997.

La *encuesta de actualización* asumió como objetivos particulares:

- Evaluar la actualización de los registros del Padrón Electoral a nivel nacional y regional.
- Estimar el nivel de algunos factores, principalmente fenómenos demográficos (cambio de domicilio, mortalidad, etc.), que inciden en la calidad del Padrón Electoral.
- Estimar la magnitud de algunas inconsistencias de los datos de registros del Padrón Electoral.

Mientras que la *encuesta de cobertura y vigencia* tuvo como objetivos:

- Conocer el grado de cobertura del Padrón Electoral y de la Credencial para Votar con fotografía, a nivel nacional y regional.
- Estimar el porcentaje de ciudadanos empadronados con domicilio de registro en la misma sección electoral en la que residen, a nivel nacional y regional.
- Estimar el porcentaje de empadronados que tienen Credencial para Votar con domicilio en la misma sección electoral en la que residen, a nivel nacional y regional.

1.5.3.2. Diseño muestral.

Inicialmente, se propuso un diseño muestral similar al de la Verificación de 1997, el marco de muestreo serían unidades geo-electorales (secciones), para finalizar con viviendas, en las cuales se preguntaría la situación en padrón de todos los ciudadanos residentes; en una segunda visita, ya comparada esta información con el padrón, se investigaría la situación de los ciudadanos omitidos. Con este procedimiento se evaluaría tanto cobertura, como actualización del padrón.

Sin embargo, al final se acordó que la Verificación del 2000 se efectuaría a partir de dos muestras independientes con un marco muestral de primera etapa común pero con unidades de observación diferentes. El esquema de muestreo quedó determinado por los objetivos de la encuesta y el recurso financiero destinado a este proyecto; así como el alcance geográfico y la precisión estadística de los indicadores. Considerando cada uno de estos aspectos se fijó un tamaño de muestra de 300 secciones. Para la encuesta de actualización se estableció que en cada sección se verificarían los datos de 50 registros en padrón.

Por otro lado, para la encuesta de cobertura y vigencia se acordó que en cada sección se llevaría a cabo un proceso de selección de unidades de muestreo secundarias, conformadas por manzanas en secciones urbanas y localidades en secciones no urbanas, concluyendo con una muestra de viviendas, que fue la tercera unidad de muestreo; para indagar en estas últimas la condición de empadronamiento de todos los ciudadanos residentes[4].

[4] *ciudadanos residentes* son todas las personas que al momento de la entrevista tenían 18 años o más, o bien que antes del 3 de julio del 2000 cumplirían esta edad y que el informante consideró que eran residentes habituales.

1.5.3.3. Regionalización.

Atendiendo a la necesidad de contar con información que reflejara las diferencias al interior del país en cuanto al grado de cobertura y nivel de actualización del padrón, se propuso una *regionalización*[5] del territorio nacional que clasificó a las 32 entidades federativas en 10 grupos. Por las características de las entidades que se agruparon, se consideró que la propuesta más conveniente sería la elaborada por CONAPO, basada en el indicador de marginación [6], y se obtuvieron las siguientes regiones:

1.- Noroeste I: Baja California, Baja California Sur y Sonora.

2.- Noroeste II: Sinaloa y Nayarit.

3.- Norte: Chihuahua, Coahuila, Nuevo León y Tamaulipas.

4.- Norte centro: Durango, San Luis Potosí y Zacatecas.

5.- Occidente: Aguascalientes, Colima y Jalisco.

6.- Centro: Guanajuato, Michoacán y Querétaro.

7.- Metropolitana: Distrito Federal, México y Morelos.

8.- Oriente: Hidalgo, Puebla, Tlaxcala y Veracruz.

9.- Sur: Chiapas, Guerrero y Oaxaca.

10.- Península: Campeche, Quintana Roo, Tabasco y Yucatán.

Para obtener estimaciones de alta confiabilidad estadística, por cada región:

- Se implementó un muestreo estratificado y polietápico,
- Se planteó alcanzar una precisión de alrededor de 3 % en el indicador de cobertura del Padrón Electoral estableciendo una confianza del 95 %.

[5] IFE, RFE, "Verificación Nacional Muestral al Padrón Electoral del año 2000, Propuesta de regionalización". Diciembre, 1999

[6] CONAPO, Desigualdad Regional y Marginación Municipal en México, 1990. México, D.F. Noviembre de 1994.

1.5.3.4. Estratificación.

Con el propósito de obtener una muestra lo más heterogénea posible en cuanto al comportamiento del padrón, en cada una de las regiones se agruparon los 2,433 municipios en 10 estratos. El método de estratificación utilizado fue el método de agrupación de las k-medias[7], el cual se aplicó conjuntamente a tres variables[8]:

- Tasa de crecimiento poblacional,
- Diferencia entre la tasa de crecimiento del padrón y de la población y
- la Cobertura del padrón.

En forma previa a la realización del operativo de campo, se realizó la actualización cartográfica correspondiente a las secciones involucradas en el universo muestral. Con esto, se sentaron las bases para obtener un marco actualizado, con un trato diferente para los casos de secciones urbanas y rurales. Es importante señalar que la actualización cartográfica se mide en dos momentos diferentes con el fin de establecer un comparativo entre ambos. A continuación se presentan los resultados obtenidos de los recorridos cartográficos.

SECCIONES URBANAS

	1er. recorrido	2do. recorrido
Total de manzanas:	8347	8901

SECCIONES RURALES

	1er. recorrido	2do. recorrido
Total de manzanas:	1253	1242

[7] Descrito en el Capítulo 3, correspondiente a Técnicas Multivariadas.

[8] IFE, RFE,"Verificación Nacional Muestral al Padrón Electoral del año 2000, Propuesta de estratificación". Diciembre, 1999.

1.5.3.5. Primera etapa de muestreo

Dividido el país en 10 regiones y 10 estratos, de las 63,634 secciones en el padrón se seleccionó una muestra de 300 secciones, la cual debía estar repartida de tal manera que se pudiera inferir el nivel de cobertura y actualización del padrón en cada región, para lo cual se fijó un tamaño de muestra de 30 secciones por región bajo las siguientes condiciones:

1. Para obtener información de todos los estratos en todas las regiones, se estableció que en cada estrato se deberan seleccionar dos secciones como mínimo,

2. Una vez que se fijaron las dos secciones por estrato, las restantes fueron distribuidas proporcionalmente en función del número total de secciones en el estrato.

Las secciones quedaron elegidas en términos de la región - estrato a la que pertenecían; cada una se seleccionó con probabilidad proporcional al tamaño (PPT), donde el tamaño de la sección se refiere a la población de 18 años y más estimada al 26 de octubre de 1999[9].

A continuación se presenta el número de secciones en muestra por región, según estrato:

[9] Con base en información censal (Censo del 90 y Conteo del 95) y suponiendo un crecimiento geométrico se estimó la población ciudadana a nivel municipal; posteriormente, la población estimada de cada municipio se distribuyó entre sus secciones electorales de acuerdo con el padrón. IFE, RFE, "Verificación Nacional Muestral al Padrón Electoral del año 2000, Selección de secciones". 5 de enero del 2000.

Región	Total Secciones	Estrato									
		1	2	3	4	5	6	7	8	9	10
1	30	-	2	2	2	3	2	6	3	8	2
2	30	-	-	-	2	3	2	9	6	6	2
3	30	2	2	2	2	3	2	7	5	3	2
4	30	2	3	-	3	5	2	7	5	3	-
5	30	2	2	-	2	7	2	5	3	4	3
6	30	2	2	-	3	6	-	10	3	2	2
7	30	-	3	-	2	7	-	6	4	5	3
8	30	2	2	2	3	2	3	4	6	4	2
9	30	2	3	2	4	3	3	3	5	3	2
10	30	-	2	2	2	2	2	8	5	4	3
Total	300										

Es importante mencionar que para fines operativos y de diseño muestral, las secciones fueron clasificadas en **urbanas** y **no urbanas**

- Por un lado, de acuerdo con la información de la base de datos del padrón, que fue el sustento de la encuesta de actualización, las 300 secciones se clasificaron, en 179 urbanas y 121 no urbanas.
- Por otro lado, la encuesta de cobertura y vigencia tuvo como referencia los productos cartográficos, con lo que, la muestra de las mismas 300 secciones quedó conformada por 181 urbanas y 119 no urbanas.

Después de la selección de secciones, las dos encuestas tuvieron unidades de observación independientes y un proceso de selección distinto, por lo cual, las etapas secundarias de muestreo se presentan por separado.

1.5.3.6. Etapas secundarias de muestreo.

- **Encuesta de actualización**[10]**.**

  Segunda y tercera etapas de muestreo.

  Debido a que en las secciones *no urbanas* la extensión territorial y la dispersión de los poblados es mayor que en las secciones urbanas, se habilitó una etapa de muestreo intermedia entre la selección de secciones y la de registros del padrón.

  Secciones urbanas

  De cada una de las 179 secciones urbanas y tomando en cuenta la base de datos del padrón correspondiente, se obtuvo una muestra de 50 registros, mediante un muestreo sistemático con igual probabilidad. Para garantizar que se seleccionarían ciudadanos de todas las edades, el ordenamiento de los registros se realizó en función de la fecha de nacimiento del ciudadano.

  Secciones no urbanas

  Para las secciones no urbanas se fijó un máximo de 8 localidades a visitar. En las secciones con menos de 8 localidades el procedimiento de selección fue el mismo que para las secciones urbanas. De las 121 secciones no urbanas, 30 tenían más de 8 localidades; entonces, para seleccionar las localidades que serían visitadas se ordenaron de acuerdo con el número de registros en el padrón y se seleccionaron mediante muestreo sistemático con igual probabilidad. Posteriormente, del conjunto de las 8 localidades se obtuvo una muestra de 50 registros.

[10] IFE, RFE,"Verificación Nacional Muestral al Padrón Electoral del año 2000, segunda y tercera etapas de muestreo para la muestra de actualización". 18 de febrero del 2000.

- **Encuesta de cobertura y vigencia [11].**

  Segunda etapa de muestreo

  A diferencia de la encuesta de actualización en que las siguientes etapas de muestreo partieron de la información en la base de datos del Padrón Electoral, en esta encuesta las siguientes etapas de muestreo se basaron en la identificación de unidades geográficas, como localidades, manzanas y grupos de viviendas (pseudomanzanas). Con el fin de tener un marco muestral de segunda y tercera etapa actualizado, fue necesario realizar un recorrido cartográfico en las secciones no urbanas y dos recorridos en las secciones urbanas. Debido a que en las secciones no urbanas la extensión territorial y la dispersión de los poblados es mayor que en las secciones urbanas, para las primeras se habilitó una etapa de muestreo intermedia entre la selección de secciones y la de manzanas.

  Secciones urbanas

  En cada una de las 181 secciones urbanas se eligieron 6 manzanas mediante un muestreo sistemático con igual probabilidad. Las manzanas se ordenaron de acuerdo con el número aproximado de viviendas, cifra que se obtuvo del primer recorrido cartográfico.

  Secciones no urbanas

  Para las secciones no urbanas se fijó un máximo de 4 localidades a visitar. En las secciones con más de 4 localidades, se eligieron las localidades que serían visitadas con probabilidad proporcional al tamaño (ppt), donde el tamaño de las localidades quedó definido por el número de manzanas.

[11] IFE, RFE."VNM2000. Proyecto general de diseño muestral. II y III etapas de la encuesta de cobertura y vigencia". 1 de marzo del 2000.

Tercera etapa de muestreo

Secciones urbanas

En cada manzana seleccionada en la segunda etapa de muestreo, se eligieron 5 viviendas mediante muestreo sistemático: el ordenamiento de las viviendas para obtener la muestra se hizo a través del consecutivo de vivienda que se asignó en el segundo recorrido cartográfico, donde se contabilizaron de manera más precisa las viviendas iniciando en la esquina noroeste de la manzana y continuando el recorrido en el sentido de las manecillas del reloj. En cada vivienda seleccionada serían censados los ciudadanos residentes.

Secciones no urbanas

En cada una de las localidades seleccionadas en las secciones no urbanas se eligieron dos manzanas o pseudomanzanas, con un muestreo aleatorio simple. En todas las viviendas habitadas de las manzanas y pseudomanzanas seleccionadas serían censados los ciudadanos residentes.

Finalmente, la muestra para cada una de las encuestas quedó comprendida por 300 secciones, 30 por región con las siguientes características:

Encuesta de actualización	Encuesta de cobertura y vigencia
179 secciones urbanas	181 secciones urbanas
121 secciones no urbanas	119 secciones no urbanas
14 909 registros	9 571 viviendas

1.5.3.7. Fórmulas y definiciones de los indicadores.

En la VNM2000, la evaluación de la calidad del padrón se cuantificó por medio de nueve indicadores: cinco básicos (con inferencia nacional y regional) y cuatro indicadores secundarios (sólo con inferencia nacional). La definición de los indicadores se apegó a la acordada para la Verificación de 1997. Adicionalmente

y a petición de los partidos políticos, se agregó un indicador que se denomina "Actualización de domicilio" con inferencia a nivel nacional y regional.

Cabe mencionar que el cálculo de los indicadores se realizó sobre los registros que presentan la característica de interés o sobre aquellos en donde se constató que se ubican en una de las condiciones complementarias, es decir, se excluyen los casos que no especificaron su situación (las no respuestas). Las fórmulas y definiciones de los indicadores se encuentran en el Apéndice C.

1.5.3.8. Resultados a nivel regional.

Las regiones 1, 3, 5, 7, 9 y 10 mostraron un nivel de actualización inferior al promedio nacional, esta situación pudo deberse a una alta movilidad de la población, estas regiones concentran a los estados fronterizos, o bien contienen un área metropolitana importante. En particular destacó la región 1 como de menor actualización, ya que era la de mayor tránsito de migración internacional. Las regiones 2, 3, 5, 7, 8 y 10 presentaron un nivel de cobertura del padrón superior al del país en su conjunto. También destacó la región 7 que contiene al Distrito Federal, Estado de México y Morelos, que tenían la mayor cobertura entre todas las regiones del país.

Adicionalmente se realizaron dos encuestas las cuales complementaron la revisión al padrón electoral 2000, a decir, "Encuesta de Retención de Credencial" y "Encuesta de Estudios de Casos".

1. **Encuesta de Retención de Credencial.**

   - **Objetivo**. Conocer qué institución y/o persona recogió la credencial

para votar y con qué fines, la fecha en que ésta se recogió al ciudadano y la fecha de la devolución.

La encuesta se levantó a todos los ciudadanos que indicaron no tener la credencial para votar con fotografía por que se las recogieron y se realizó con el ciudadano en cuestión (o un informante) y la información resultante fue avalada por el nombre, firma o huella del mismo.

- **Resultados**. Las cédulas levantadas por el operativo de campo fueron 33, de las cuales, en 26 encuestas la información la proporcionó el ciudadano en cuestión (86.7 %), 4 por un informante familiar (13.3 %), y 3 quedaron pendientes. El 53.3 % fue recogida por una institución gubernamental principalmente para cuestiones de salud y asistencia social, el 43.3 % por instituciones particulares para la atención de asuntos personales y administrativos.

2. **Encuesta de Estudios de Casos.**

- **Objetivo**[12]. Constatar la existencia de los ciudadanos cuyo registro electoral tiene alguna de las dos siguientes características:

  a) Ciudadanos con "domicilio conocido", que agrupó a la población que al momento de manifestar sus datos para quedar inscritos en el Padrón Electoral informaron no contar con una nomenclatura (nombre de la calle, colonia, número exterior y número interior) para identificar su lugar de residencia. Los domicilios conocidos estaban ubicados principalmente en zonas rurales donde la población es baja, lo que hizo posible identificar a los ciudadanos.

  b) Ciudadanos "indígenas". Dado que no se contaba con la identificación de la población indígena, se tomó como base la información del

[12]IFE, RFE "VNM 2000 Estudios de caso. Diseño Muestral", 6 abril de 2000.

Instituto Nacional de Estadística, Geografía e Informática (INEGI) del "Conteo de Población y Vivienda 1995", y así definir como indígenas a los habitantes de los municipios con predominancia en habla de alguna lengua indígena.

Después de la Verificación Nacional de 1997, la VNM2000 tuvo dos antecedentes: el Programa de Diagnóstico - Mejoramiento y en su caso Corrección del Padrón Electoral (PRODIMEC, 1998); y la Técnica Censal Parcial aplicada en los municipios de Reynosa y Río Bravo en 1999. Estos dos últimos programas se propusieron evaluar y mejorar la calidad del padrón en áreas específicas. La selección de estas áreas se basó, principalmente, en métodos demográficos y los resultados sólo reflejan la situación de la zona atendida; sin embargo, con su ejecución se capitalizó la experiencia que permitió afinar los mecanismos para obtener la información.

1.5.4. Verificación Nacional Muestral 2003.

Dentro de los trabajos de preparación de las elecciones federales del año 2003, la Comisión Nacional de Vigilancia (CNV), integrada por 11 partidos políticos con registro nacional y la Dirección Ejecutiva del Registro Federal de Electores (DERFE), elaboró un estudio sobre la calidad del Padrón Electoral el cual se agrega a la serie de evaluaciones muestrales al Padrón Nacional, que habían estado presentes desde las elecciones federales de 1994. En términos generales, los resultados de esta verificación 2003 fueron comparables con los del diagnóstico de 1996 y las verificaciones de 1997, 2000 y 2002; y en este sentido, los resultados de la VNM03 además de dar a conocer la situación actual del padrón, permitieron analizar su evolución.

La experiencia adquirida con la realización de las últimas cuatro evaluaciones muestrales se aprovechó durante el desarrollo de esta verificación, tanto en la definición de objetivos, diseño conceptual y estadístico, como en la ejecución del operativo y la coordinación de tareas entre la DERFE y los partidos políticos.

Al igual que en la verificación del 2000, el diseño conceptual apuntó hacia la realización de dos encuestas: la Encuesta de Cobertura y la Encuesta de Actualización con las mismas orientaciones. En cuanto a la participación de los partidos políticos, las dos encuestas fueron supervisadas por cada una de las representaciones políticas y, de acuerdo con su experiencia, intervinieron en las etapas que cada una consideró conveniente controlar.

1.5.4.1. Objetivos.

El objetivo general de la VNM03 fué evaluar la calidad del Padrón Electoral. Concretamente, la evaluación consistió en estimar indicadores con dos fines:

1. Orientar el veredicto del Consejo General sobre la validez y suficiencia de los productos electorales que en su momento, se utilizaron en las elecciones federales 2003, y
2. Apoyar la planeación de programas de mejoramiento referentes la calidad del padrón.

La aportación más importante de esta verificación **fue la evaluación del padrón a nivel estatal**, lo cual permitió caracterizar a las entidades federativas de acuerdo a las condiciones del empadronamiento y calidad del padrón, elemento útil para la planeación de programas de mejoramiento de la calidad del mismo.

1.5.4.2. Diseño muestral.

El esquema de muestreo fue similar al de la verificación del 2000: polietápico y estratificado, con la primera etapa de muestreo común para las dos encuestas (selección de secciones) y las siguientes etapas de muestreo fueron independientes entre una y otra encuesta.

La determinación del tamaño de muestra y su distribución obedeció al interés por obtener indicadores a nivel estatal y al recurso financiero disponible. Para reducir el intervalo de variación de las estimaciones, se llevó a cabo un análisis estadístico previo a la selección de la muestra, que consistió en formar estratos de secciones al interior de cada estado[13].

1.5.4.3. Regionalización.

Dado que uno de los propósitos de esta verificación fue obtener indicadores a nivel estatal, el tamaño de muestra se debía fijar por entidad; sin embargo, la información disponible era insuficiente para decidir el tamaño de muestra por estado. La alternativa fue emplear estimaciones de las 10 regiones definidas en la VNM2000 y **suponer que el comportamiento de cada estado era similar al de la región a la que pertenecía**[14]. El procedimiento para calcular el tamaño de muestra fue el siguiente:

- Se eligieron tres indicadores sobre los que se harían los cálculos: Cobertura del Padrón Electoral Nacional, Vigencia de la Credencial para Votar y Actualización del Padrón Electoral.

[13] IFE-RFE, "VNM03. Estratificación por entidad federativa", 4 de diciembre de 2002.
[14] IFE-RFE, "Aspectos técnicos de la VNM03. 2do. documento de diseño conceptual", 21 de octubre de 2002.

- De la VNM2000 se empleó, tanto la estimación puntual de estos indicadores, como la estimación de la varianza por región ($\sigma^2_{diseño}$).
- Se estimó el efecto de diseño de la VNM2000 (Deff).

$$Deff = \frac{\sigma^2_{diseño}}{\sigma^2_{mas}}$$

- Se calculó el coeficiente de correlación intraclase (ρ_0) de las unidades últimas de muestreo dentro de las unidades primarias de muestreo.
- Suponiendo una reducción del número de unidades últimas de muestreo y con el mismo ρ_0 de la VNM2000, se estimó un nuevo *Deff**:

$$Deff^* = 1 + (m^* - 1)\rho_0$$

- Se estimó el tamaño de muestra suponiendo un muestreo aleatorio simple.
- Multiplicando el nuevo *Deff** por el tamaño de muestra (vía MAS), se calcula el tamaño de muestra del diseño propuesto para la VNM03.
- Variando las precisiones de los tres indicadores y suponiendo una confianza del 95 %, se elaboraron 10 escenarios de tamaño de muestra según la precisión que se alcanzaría en cada indicador[15,16].

Para todo el país se fijó un tamaño de muestra de 2,500 secciones y dado que el diseño de esta verificación tuvo modificaciones respecto a la del 2000, no se pudieron establecer las precisiones que alcanzarían las estimaciones nacionales. Particularmente, se obtuvo un primer tamaño de muestra por entidad federativa

[15] IFE-RFE, "VNM03. Precisiones estimadas según siete escenarios de tamaño de muestra", 28 de octubre de 2002.

[16] IFE-RFE, "VNM03. Precisiones estimadas según tres nuevos escenarios de tamaño de muestra", 6 de noviembre de 2002.

en el cual se procuró simultáneamente obtener estimaciones con una precisión alrededor del 2 % para la Cobertura del Padrón Electoral Nacional, 3.5 % para la Actualización del Padrón Electoral y 5.5 % para la Vigencia de la Credencial.

1.5.4.4. Estratificación.

Al igual que en las verificaciones del 2000 y del 2002, previo a la selección de la muestra se hizo un análisis estadístico para formar estratos de secciones. El propósito de la estratificación fue dividir cada estado, agrupando unidades geográficas lo más parecidas en cuanto a las características del empadronamiento y la calidad del padrón. En esta verificación el procedimiento general fue agrupar municipios similares y posteriormente clasificar las secciones en urbanas y no urbanas.

Para dividir el país, clasificando unidades geográficas similares, fue necesario recurrir a fuentes de información con cobertura nacional, por lo que se hizo uso del censo de población y del propio padrón. Aprovechando los resultados regionales de la VNM02 (5 regiones), se asociaron indicadores de la situación del empadronamiento (estimados con la VNM02) con variables sociodemográficas.

Las asociaciones entre indicadores de la calidad del empadronamiento y variables sociodemográficas, se evaluaron mediante el análisis de correlación canónica con el fin de seleccionar por región, las variables sociodemográficas que más se vinculan al comportamiento de los indicadores de la calidad del empadronamiento. Con las variables seleccionadas, se estratificó aplicando el método de conglomerados de Ward[17].

[17] Descrito con detalle en el Capítulo 3.

1.5.4.5. Etapas de muestreo.

Esquema de muestreo de la Encuesta de Cobertura

Etapa de muestreo	Unidades de muestreo			Método de selección
Primera	Secciones (2 500)			PPT Probabilidad proporcional a la población estimada de la sección, con reemplazo
	Urbanas	Mixtas	Rurales	
Segunda	Manzanas (5 por sección)	Manzanas (2 por sección) Localidades (4 por sección)	Localidades (6 por sección)	Manzanas: Sistemático con arranque aleatorio ordenado por número de viviendas de mayor a menor Localidades: PPT, probabilidad proporcional al padrón de la localidad, con reemplazo
Tercera	Viviendas (3 por manzana)	Viviendas (3 por manzana o localidad*)	Viviendas (3 por localidad*)	Área urbana: Sistemático con arranque aleatorio ordenado por consecutivo de vivienda de menor a mayor Area no urbana: "mas"

⋆ En las secciones con más de 500 registros en padrón se eligieron 4 viviendas, previo a la selección de viviendas se eligieron dos manzanas.

Esquema de muestreo de la Encuesta de Actualización

Etapa de muestreo	Unidades de muestreo		Método de selección
Primera	Secciones (2 500)		PPT Probabilidad Proporcional a la población estimada de la sección con reemplazo
	Urbanas	Mixtas Rurales	
Segunda	-	Localidades (6 por sección)	PPT Probabilidad Proporcional al padrón de la localidad con reemplazo
Tercera	Registros* (30 por sección)	Registros (30 por sección)	Sistemático con arranque aleatorio ordenado por edad de mayor a menor

⋆ En las secciones urbanas, la selección de registros se realizó en la segunda etapa de muestreo.

1.5.4.6. Resultados.

Los resultados se presentaron en cuatro niveles: **nacional, área urbana y no urbana, regional** (10 agrupaciones de estados geográficamente contiguos) **y estatal**. Para los primeros tres niveles, además de presentarse las estimaciones del 2003, se hizo una comparación con la situación que se registró en el 2000. Las precisiónes para los indicadores con los que se diseñó la muestra, Cobertura

del Padrón Electoral, Vigencia de la Credencial para Votar y Actualización del Padrón Electoral, en esta verificación resultaron mejores a las obtenidas en las cuatro evaluaciones muestrales anteriores.

Analizando los cambios que se presentaron de 2000 a 2003, a partir de los once indicadores que eran comparables, se confirmó que hubo mejoras en la inscripción al padrón y la solicitud de la credencial, puesto que la Cobertura del Padrón Electoral y la de la Credencial para Votar (ambos a nivel nacional) tuvieron aumentos estadísticamente significativos.

1.5.5. Verificación Nacional Muestral 2006.

Dentro de los trabajos de preparación para las elecciones federales del 2 de julio de 2006 la Comisión Nacional de Vigilancia, integrada por ocho partidos políticos con representación nacional y la Dirección Ejecutiva del Registro Federal de Electores (DERFE), desarrollaron el proyecto "Verificación Nacional Muestral 2006" (VNM06).

En cuanto al diseño conceptual, se siguió la línea de las verificaciones de 1996 a 2005. Se diseñaron dos cuestionarios: el primero para registrar datos sobre el empadronamiento de los ciudadanos y el segundo para verificar el domicilio de registro de los ciudadanos inscritos en el padrón, la residencia del ciudadano y las causas de no residencia.

1.5.5.1. Objetivos.

El objetivo general de la VNM06 fue generar indicadores para evaluar el Padrón Electoral y los productos electorales que serían empleados en las elec-

ciones federales del 2 de julio de 2006. La utilidad inmediata de los resultados fué la de proporcionar elementos de juicio al Consejo General del Instituto Federal Electoral para que emita un veredicto sobre la validez y definitividad del Padrón Electoral y Lista Nominal que serían empleados en las elecciones.

El esquema general para atender estos objetivos consistió en la realización de dos encuestas: Encuesta de Cobertura y Encuesta de Actualización. Con la primera, se midieron las condiciones de empadronamiento de los ciudadanos residentes en el país; mientras que con la encuesta de actualización se evaluaron algunos aspectos de la calidad de los registros en el padrón.

1.5.5.2. Diseño muestral.

El esquema de muestreo fue polietápico y estratificado. Puesto que debían obtenerse estimaciones para dos divisiones geográficas del país simultáneamente, (estados y categorías de distritos), previo a la distribución de la muestra se elaboró una clasificación de distritos. El tamaño de muestra y su distribución quedó determinado por el detalle de inferencia requerido (a nivel nacional, estatal y categorías de distritos).

1.5.5.3. Estratificación.

La estratificación[18] tuvo dos intenciones: mostrar diferencias en las condiciones de empadronamiento entre grupos de distritos y disminuir la varianza muestral de las estimaciones. El procedimiento se hizo en dos pasos, el primero fue la clasificación de los distritos electorales y el segundo fue la agrupación de las secciones al interior de cada distrito.

[18] IFE-DERFE, "Verificación Nacional Muestral 2006. Diseño estadístico", 17 de noviembre de 2005.

- **Clasificación de los distritos.**

El propósito de clasificar los 300 distritos electorales fue garantizar la distribución de la muestra entre todas las categorías de esta unidad geoelectoral; así, ya fue posible hacer la comparación de las estimaciones entre grupos de distritos.

Para la agrupación de los distritos se debía establecer qué características definirían su similitud y con qué variables serían evaluadas. Sin embargo, los indicadores de las verificaciones muestrales son medidas de las condiciones del empadronamiento, que no se emplearon para la estratificación porque las muestras fueron insuficientes para obtener estimaciones por distrito. Por tal motivo, se recurrió a otras fuentes de información: a) la base de datos del padrón, b) la cartografía electoral y c) censos de población.

1.5.5.4. Etapas de muestreo.

Esquema de muestreo de la Encuesta de Cobertura

Etapa de muestreo	Unidades de muestreo			Método de selección
Primera	Secciones (2 940)			PPT
	Urbanas	Mixtas	Rurales	Probabilidad proporcional a la población estimada de la sección, con reemplazo.
Segunda	5 manzanas por sección	3 manzanas por sección 3 localidades por sección	6 localidades por sección	Manzanas: Sistemático con arranque aleatorio ordenado por número de viviendas de mayor a menor. Localidades: PPT, probabilidad proporcional al padrón de la localidad, con reemplazo.
Tercera	3 viviendas por mz	3 viviendas por mz	-	Muestreo aleatorio simple
	-	En localidades con menos de 500 registros en padrón **3 viviendas por localidad** En localidades con 500 o más registros en padrón **2 manzanas por localidad**		Muestreo aleatorio simple
Cuarta	-	En localidades de 500 o más registros en padrón **2 viviendas por manzana**		Muestreo aleatorio simple

Esquema de muestreo de la Encuesta de Actualización

Etapa de muestreo	Unidades de muestreo	Método de selección
Primera	Secciones (2 490)	PPT Probabilidad proporcional a la población estimada de la sección con reemplazo
Segunda	Registros (30 por sección)	Sistemático con arranque aleatorio ordenado por edad de mayor a menor

1. **Primera etapa de muestreo**[19]

   El tamaño de muestra de la primera etapa se fijó en 2,940 secciones. El método de muestreo para las secciones fue estratificado donde cada sección se seleccionó con probabilidad proporcional a su población estimada. En los estratos donde el número de secciones por seleccionar era igual al total de secciones en el estrato, se seleccionaron todas las secciones. Dado que el proceso de selección fue con reemplazo, la muestra que se obtuvo contenía secciones repetidas.

2. **Etapas secundarias de muestreo**

   La verificación fue diseñada para evaluar el padrón desde dos perspectivas, y cada enfoque tiene su población de estudio, por un lado, para la Encuesta de Cobertura, la población de estudio son los ciudadanos residentes en el estado; en tanto que, para la Encuesta de Actualización, la población de estudio son los registros en el padrón y la lista nominal. Por ello, después de la primera etapa de selección, cada encuesta tuvo su propio esquema de muestreo.

   I.- ENCUESTA DE COBERTURA. Previo a la selección de manzanas se efectuó un operativo de campo para actualizar la cartografía de las secciones en muestra. Durante este operativo, además de actualizar los planos cartográficos, se elaboró un lista con el número aproximado de viviendas habitadas por manzana, el cual sirvió para definir el marco de muestreo.

   La muestra de manzanas quedaría conformada sólo por aquéllas que tuvieran al menos una vivienda habitada. El procedimiento para seleccionar las

[19] IFE-DERFE, "Verificación Nacional Muestral 2006. Selección de secciones", 24 de noviembre de 2005.

manzanas fue sistemático con arranque aleatorio, las manzanas con viviendas habitadas de cada sección se ordenaron en forma ascendente respecto al número de viviendas habitadas.

a) En las secciones urbanas se seleccionaron cinco manzanas, si la sección tenía menos de cinco manzanas se seleccionaron todas.
b) En las secciones mixtas se seleccionaron tres manzanas, si la sección tenía menos de tres manzanas se seleccionaron todas. Para cada sección se hizo la selección de manzanas tantas veces como estuvo repetida la sección en la muestra. De este procedimiento resultaron seleccionadas 10,310 manzanas.

Selección de localidades. Para construir el marco de muestreo se clasificaron las localidades en tres grupos:

1) Localidades con soporte cartográfico[20] y registros en el padrón[21] (7,097).
2) Localidades con soporte cartográfico que no tienen registros en padrón (4,517). A estas localidades se les asignó un empadronado, para que tengan probabilidad de selección.
3) Localidades sin soporte cartográfico ni registros en el padrón (122).

El marco de muestreo quedó integrado por las localidades del grupo 1) y 2), en total 11,614 localidades. Del grupo 3) no se incluyó localidad alguna, sólo se utilizó el análisis para adecuar las probabilidades de selección de las demás localidades. El método de selección de localidades fue con probabilidad proporcional al padrón de la localidad, con las siguientes características:

[20] Catálogo de localidades con soporte cartográfico al 6 de febrero de 2006.
[21] Corte de padrón al 15 de enero de 2006.

a) En las secciones mixtas se seleccionaron tres localidades, si la sección tenían menos de tres localidades se seleccionaron todas.
b) En las secciones rurales se seleccionaron seis localidades, si la sección tenía menos de seis localidades se seleccionaron todas.

Si la sección estaba en muestra más de una vez, la selección de localidades al interior de ella se realizó tantas veces como se repitió la sección. De este procedimiento resultaron seleccionadas 2,731 localidades.

Selección de viviendas[22]. La selección de viviendas se realizó mediante muestreo sistemático con arranque aleatorio; el orden de las viviendas se estableció de acuerdo al consecutivo dentro de cada manzana. Para establecer cuál vivienda era la de reemplazo, en las manzanas seleccionadas con cuatro viviendas habitadas se generó un número aleatorio entre cero y uno para cada vivienda en la manzana; la vivienda con el mayor número aleatorio fue la de reemplazo. Como resultado de este procedimiento quedaron seleccionadas 39,092 viviendas, de las cuales 29,797 eran viviendas para ser visitadas de manera obligatoria y 9,295 eran de reemplazo.

II.- ENCUESTA DE ACTUALIZACIÓN. El método de selección de registros fue sistemático con arranque aleatorio. En todas las secciones se seleccionaron 30 registros y se replicó el proceso en las secciones repetidas según el número de veces que estaban en la muestra. Para la selección, los registros de cada sección se ordenaron simultáneamente con dos criterios, el primero respecto a su fecha de nacimiento (ascendente), y el segundo res-

[22] IFE-RFE, "Verificación Nacional Muestral 2006. Encuesta de Cobertura. Selección de viviendas en secciones urbanas y mixtas"

pecto a la fecha del último trámite realizado (descendente).

De este procedimiento resultaron seleccionados 88,200 registros; sin embargo, en las secciones repetidas 20 registros fueron seleccionados dos veces, por lo que la muestra efectiva fue de 88,180 registros[23].

1.5.5.5. Resultados.

Para cada encuesta, los resultados se presentaron en cuatro niveles: 1) nacional, 2) entidad federativa, 3) tipo de distrito y 4) tipo de sección (urbanas y no urbanas). Asímismo se presentó la comparación de los resultados nacionales y estatales con las estimaciones de la VNM2005. Tanto el cotejo de las estimaciones estatales como los resultados de las pruebas estadísticas para la comparación de estos datos, se encuentra en IFE - DERFE, "Verificación Nacional Muestral 2006. Informe de resultados.", 2 de mayo de 2006.

Verificaciones Muestrales realizadas en años pre - electorales

1.5.6. Verificación Nacional Muestral 1996.

La Comisión Nacional de Vigilancia en su acuerdo 2-70 del 14 de diciembre de 1995 aprobó la realización de un estudio de diagnóstico muestral del Padrón Electoral con el propósito de "estimar la cobertura respecto de la población de 18 años y más, e identificar el nivel de desactualización". Esta actividad se desarrolló durante los primeros meses de 1996.

[23] DERFE-IFE, "Verificación Nacional Muestral 2006. Encuesta de Actualización. Selección de registros", 26 de enero de 2006.

El marco muestral se integró con las secciones electorales, el método de selección fue por conglomerados de una etapa con selección sistemática de secciones; el nivel de confianza se fijó en 95 % y la precisión de 3 %. Los resultados se presentaron en dos vertientes:

Cobertura

Cobertura del padrón: 88.7 %
Vigencia del padrón: 71.3 %
Cobertura de la credencial: 80.8 %
Vigencia de la credencial: 66.9 %

Actualización

Desactualización del padrón: 21.5 %
Rezago del padrón: 28.7 %

1.5.7. Verificación Nacional Muestral 2002.

La Verificación Nacional Muestral 2002 (VNM02) fue una encuesta para evaluar el estado del empadronamiento electoral del país. La propuesta de evaluación incluyó generar datos para el diseño de mejores estrategias de empadronamiento, de credencialización, y en particular para la ubicación de módulos de atención ciudadana. En este sentido, esta encuesta significó una estrategia innovadora sobre la producción de datos útiles para los planes del Registro Federal de Electores. Para tal fin se hizo una recolección de datos por medio de una encuesta probabilística la cual incorpora en su trabajo de campo, la realización de entrevistas directas con los ciudadanos en el mismo domicilio donde tienen su residencia. Se realizaron preguntas sobre la situación electoral de cada uno de los ciudadanos residentes en las viviendas.

La "Verificación Nacional Muestral 2002", tuvo un desarrollo notable con la particular intención de proveer información de mayor utilidad para la planeación del empadronamiento. Su pretensión más ambiciosa fué la de proporcionar estimaciones de la calidad del empadronamiento en los municipios del país, reto que requirió de un planteamiento audaz en los objetivos y en el diseño estadístico de la encuesta.

Con la VNM02 se fortalecieron notablemente las estimaciones a nivel municipal (pero estas estimaciones no fueron directas); y con ello, los datos para la planeación fueron mejorados. Además se abrió la posibilidad de analizar con mayor profundidad algunas de las asociaciones de la calidad del empadronamiento con aspectos socioeconómicos y demográficos de la población.

1.5.7.1. Regionalización.

Con la intención de elevar el nivel de estimación, se consideró conveniente que, previo a la selección de unidades en muestra, se dividiera al país en cinco estratos o regiones conformadas por grupos de municipios o estados[24]. **Con ello, el propósito fué construir un modelo de regresión para cada región, con lo que se tuvo esperanza de obtener mejores estimaciones con modelos regionales que aquellos que se obtendrían con un solo modelo para todo el país.** Asimismo, con estas regiones se esperaba aumentar la capacidad de predicción de dichos modelos de regresión.

Para elegir la división territorial más eficiente a los propósitos de la encuesta se estudiaron varios escenarios de estratificación. Este análisis incluyó:

[24] IFE-RFE, "Verificación Nacional Muestral del 2002. Diseño estadístico (primer avance)", 26 de septiembre de 2001.

- Definición y construcción de un marco de indicadores.
- Evaluación de escenarios de estratificación del país.

La división del país se estableció a partir del análisis de varios escenarios de agrupaciones de estados y municipios, y consistió en evaluar las posibilidades de predicción de modelos de regresión en cada grupo de estados y municipios. El criterio para elegir la regionalización más apropiada fue seleccionar las cinco regiones cuyos modelos de regresión, en donde la variable "relación padrón-censo" (RPC: población-padrón/población)[25] fuera explicada por variables censales y tuvieran la mayor R^2 (proporción de variabilidad que las variables censales explican del indicador RPC).

De estos ejercicios se obtuvo que la agrupación más adecuada es la división del país en las siguientes cinco regiones de estados:

Región	**Estados**
I	Baja California, Baja California Sur, Coahuila, Chihuahua, Nuevo León, Sonora y Tamaulipas.
II	Durango, Nayarit, San Luis Potosí, Sinaloa y Zacatecas.
III	Aguascalientes, Colima, Guanajuato, Jalisco, Michoacán y Querétaro.
IV	Distrito Federal, Hidalgo, Estado de México, Morelos, Puebla, Tlaxcala y Veracruz.
V	Campeche, Chiapas, Guerrero, Oaxaca, Quintana Roo, Tabasco y Yucatán.

[25] Este indicador fue construido con información del conteo de 95. PROGRESA-CONAPO, Índices de marginación, 1995, México, D.F., diciembre de 1998.

1.5.7.2. Objetivos.

El proyecto de la Verificación Nacional Muestral 2002 se desarrolló a partir de dos objetivos generales:

- Para el país en su conjunto, proporcionar indicadores para evaluar las condiciones de empadronamiento,
- Para unidades geográficas más detalladas, generar información que oriente y dé sustento a las acciones encaminadas a elevar el nivel y la vigencia del empadronamiento[26].

Aprovechando la experiencia de los trabajos de verificación y diagnóstico del padrón realizados por el Registro, se consideró que una forma de atender los objetivos planteados era realizar una encuesta donde la población de estudio fuera los ciudadanos residentes en el país. El procedimiento operativo consistiría en indagar, a través de una entrevista con los ciudadanos en su domicilio de residencia, su situación de registro en el padrón, la vigencia de sus datos de registro y tenencia de la credencial para votar.

Después de conciliar los requerimientos de información y determinar el nivel geográfico de inferencia que se podría alcanzar, se establecieron los siguientes objetivos específicos:

- A nivel nacional y para las cinco regiones en que se dividió el país, obtener indicadores para evaluar las condiciones del empadronamiento, comparables con los de la encuesta de cobertura de la VNM2000, siendo éstos:

  1. Cobertura del Padrón Electoral.

[26] IFE-RFE, "Encuesta para evaluar el empadronamiento en el año 2002", 9 de julio de 2001.

2. Vigencia del Padrón Electoral.
3. Cobertura de la Credencial para Votar.
4. Vigencia de la Credencial para Votar.

- A nivel nacional y para las cinco regiones, inferir la proporción de ciudadanos que no tienen credencial del domicilio donde residen; o bien, si su credencial tuvo algún error en sus datos y no habían solicitado una nueva credencial, indicador denominado *demanda potencial de solicitudes de credencial*.
- Estimar la demanda potencial de solicitudes de credencial a nivel municipal, con el fin de apoyar la planeación de la campaña de actualización previa a las elecciones federales del 2003.

1.5.7.3. Diseño muestral.

El diseño muestral se determinó en función de dos criterios:

1. Se debía atender los objetivos establecidos con alta precisión estadística, y
2. Adaptarse a los recursos financieros y tiempo disponible para este proyecto.

La población de estudio de la encuesta fueron los ciudadanos residentes en el país. Para obtener una muestra de ellos, de forma que se conociera su probabilidad de selección y con ello inferir las caractersticas de la población, fue necesario seleccionar la muestra en varias etapas[27] .

- **Primera etapa de muestreo**. En cada región se seleccionaron 80 municipios para un total de 400 municipios en muestra. La selección de los municipios se hizo con muestreo sistemático, el orden que se dió fue el valor

[27] IFE-RFE, VNM02. "Diseño muestral (avance)", 24 de noviembre de 2001.

del indicador RPC. Este tipo de selección permite generar las condiciones necesarias para la aplicación de análisis estadísticos.

- **Segunda etapa de muestreo**. Al interior de los municipios en muestra se seleccionaron seis secciones con probabilidad proporcional al número estimado de ciudadanos residentes en la sección y con reemplazo (ppt). De esta selección se esperaba obtener una muestra 2 mil 400 secciones; sin embargo, debido a que algunos municipios tienen menos de seis secciones y que algunas secciones salieron más de una vez seleccionadas, se obtuvo una muestra de mil 901 secciones.

- **Tercera etapa de muestreo**. En esta etapa de selección hubo dos unidades de muestreo: **manzanas y localidades**, dependiendo del grado de urbanización de las secciones. En las secciones urbanas se eligieron manzanas, en las secciones rurales se eligieron localidades y en las secciones mixtas manzanas y localidades.

- **Cuarta etapa de muestreo**. La selección de viviendas se hizo con dos procedimientos, la forma de elegir viviendas dependió de la unidad de muestreo que fue seleccionada en la etapa anterior.

    1. **Selección de viviendas en manzanas**.

       Una vez que se obtuvo la muestra de manzanas en las secciones urbanas y en la parte amanzanada de las secciones mixtas, se procedió a hacer un segundo recorrido cartográfico para obtener el número exacto de viviendas incluyendo la identificación de cada una de ellas. Se seleccionaron dos viviendas por manzana y cuando la manzana tenía dos o menos viviendas se tomaron todas. La selección fue por muestreo sistemático, ordenadas de forma ascendente por el número consecutivo que le fue

asignado en el segundo recorrido cartográfico.

2. **Selección de viviendas en localidades rurales**.
   En las localidades elegidas en las secciones rurales y mixtas, se seleccionaron tres viviendas por localidad. En caso de que la localidad tuviera tres o menos viviendas se eligieron todas. El método de selección fue aleatorio simple sin reemplazo(MAS). El procedimiento de selección fue realizado por el visitador domiciliario, quien previamente recorrió la localidad para contabilizar el total de viviendas habitadas y asignar un número consecutivo a cada una de las viviendas, empleando una tabla de números aleatorios por localidad[28].

1.5.7.4. Análisis estadístico.

Debido a que la estimación de la demanda potencial de solicitudes de credencial a nivel municipal fue el objetivo que estableció mayores exigencias al tamaño y distribución de la muestra, el análisis estadístico se centró en la elección y aplicación de métodos multivariados que optimizaran el alcance de los datos muestrales.

El procedimiento elegido se basó en el uso de modelos de regresión predictivos; con ello el propósito fue asociar indicadores de la situación de empadronamiento (estimados con la encuesta) y variables conocidas para todos los municipios (provenientes de otras fuentes de información); de esta forma no fue necesario dispersar las observaciones en todos y cada uno de los municipios del país, bastó con distribuir la muestra en algunos municipios, con la condición de

[28] IFE-RFE, "Verificación Nacional Muestral, 2002. Manual del Técnico Cartógrafo", pp. 17-23, Mayo 2002.

que las características conocidas de los municipios seleccionados reflejaran variabilidad del indicador a estimar. El procedimiento fue el siguiente:

1. Obtener una muestra de municipios con diversos perfiles sociodemográficos. Las características sociodemográficas de los municipios se midieron a través de indicadores calculados con información censal y Padrón Electoral.

2. Con base en los datos muestrales, inferir el nivel del indicador *demanda potencial de solicitudes de credencial* en cada uno de los municipios en muestra.

3. Construir modelos de regresión para establecer la relación que hay entre el indicador *demanda potencial de solicitudes de credencial* y las características sociodemográficas de los municipios en muestra.

4. Una vez establecido el modelo de regresión y comprobada su capacidad predictiva, aplicar el modelo a todos los municipios, incluso para los que no quedaron seleccionados en la muestra (imputación).

La aplicación de este procedimiento apoyado en información censal, tiene como ventaja producir estimaciones de gran utilidad para la planeación de programas de empadronamiento; por un lado, la especificidad de los conceptos medidos con la encuesta; y por otro lado, la cobertura y detalle geográfico del censo permiten estimar para cada municipio la demanda potencial de credenciales.

1.5.7.5. Definición de indicadores y fórmulas para su obtención.

Apegado a los objetivos de la encuesta, el cuestionario se diseñó para obtener información que permitiera estimar cinco indicadores. La definición de cuatro de los indicadores corresponde a la acordada para la verificación del 2000. El cálculo de los indicadores se hace sobre los registros que presentan la característica de

interés o se puede decir que queda bien ubicado en las categorías complementarias; esto es que, se excluyen los casos que no especificaron su situación (las no respuestas). La definición de indicadores así como las fórmulas para su obtención se presentan en Apéndice C.

1.5.7.6. Análisis de resultados a nivel nacional.

En la comparación de los resultados de la VNM02 con los resultados de la VNM2000, a nivel nacional, se pudo observar:

- La cobertura del padrón tuvo un ligero aumento, 0.26 puntos porcentuales; sin embargo, no se tiene suficiente evidencia estadística para decir que el incremento es significativo.
- La vigencia del padrón tuvo un ligero aumento de 0.09 puntos porcentuales.
- La cobertura de la credencial para votar tuvo una ligera disminución, 0.43 puntos porcentuales.
- La vigencia de la credencial para votar tuvo una ligera disminución, 0.84 puntos porcentuales.

Para los tres últimos puntos se concluyó que los cambios no fueron estadísticamente significativos.

1.5.7.7. Análisis de resultados a nivel regional.

El aumento más notable en el nivel de los cuatro indicadores se dió en la Región I; el incremento en la cobertura del padrón (3.34 puntos porcentuales) y de

la credencial (3.33 puntos porcentuales) fue significativo. Sin embargo, los incrementos en la vigencia del padrón (6.97 puntos porcentuales) y de la credencial (6.47 puntos porcentuales) no fueron significativos.

En cuanto a la disminución de los indicadores, los cambios más notables son los de la Región II; las reducciones en la vigencia del padrón (7.85 puntos porcentuales) y de la credencial (9.58 puntos porcentuales) fueron estadísticamente significativas. En tanto que las disminuciones en la cobertura del padrón (1.31 puntos porcentuales) y de la credencial (2.28 puntos porcentuales) no resultaron estadísticamente significativas.

Finalmente, aún cuando al interior del país, para cada indicador hubo una región que presentó cambios significativos, éste no se reflejó en el comportamiento nacional porque otras regiones apuntaron en sentido opuesto.

1.5.8. Verificación Nacional Muestral 2005.

Durante los meses de febrero y marzo de 2005, en sesiones ordinarias, extraordinarias y mesas de trabajo de CONASE se abordó el tema de la Verificación Nacional Muestral 2005; la discusión se centró en los objetivos y el diseño general. Los representantes de los partidos políticos manifestaron su interés por un **estudio con inferencia nacional, pero sobre todo con inferencia estatal**. En cuanto a la temática de las encuestas, éstos solicitaron que se obtuviera información sobre dos aspectos:

a) de los ciudadanos que viven en un domicilio distinto al domicilio de empadronamiento y

b) causas por las que los ciudadanos que viven en el domicilio de empadronamiento no están presentes al momento de la entrevista.

Bajo estas consideraciones, el 7 de abril de 2005, se aprobó el esquema general de la Verificación Nacional Muestral 2005; así como el cronograma de actividades. Además, en el mismo acuerdo se estableció que el levantamiento de información para los estados de México y Nayarit se haría después de sus procesos electorales locales[29]. A partir de esta aprobación se desarrollaron las actividades de: diseño conceptual, estadístico y planeación operativa; así como, levantamiento, captura y procesamiento de la información.

1.5.8.1. Objetivos

Fueron dos los objetivos generales de la VNM05:

- Conocer las condiciones del empadronamiento y la situación de los registros en la base de datos del padrón electoral y
- Contar con indicadores que apoyen la planeación de la campaña anual intensa previa a las elecciones federales de 2006.

El esquema general para atender los objetivos fue el mismo que se siguió en Verificaciones anteriores: la realización de las Encuestas de Cobertura y de Actualización. Los objetivos específicos de la Encuesta de Cobertura fueron:

- Conocer el nivel de empadronamiento de los ciudadanos residentes en el país.
- Estimar la proporción de ciudadanos que tienen Credencial para Votar.
- Estimar la proporción de empadronados que viven en la sección electoral de registro.

[29] Acuerdo del Comité Nacional de Supervisión y Evaluación O-008-03 del 7de abril de 2005

- Estimar la demanda potencial de solicitudes de Credencial para Votar para apoyar la distribución de módulos de la campaña de actualización previa a las próximas elecciones federales.

Los objetivos específicos de la encuesta de Actualización fueron:

- Constatar que cada registro del Padrón Electoral y de la Lista nominal corresponda a un ciudadano.
- Medir el impacto de factores demográficos (migración y mortalidad) en la calidad del Padrón electoral y la Lista nominal.

1.5.8.2. Diseño Muestral.

El esquema de muestreo fue polietápico y estratificado. La determinación del tamaño de muestra y su distribución estuvo en función de:

1. El nivel de inferencia requerido para las estimaciones y
2. El presupuesto disponible.

Dado el interés por alcanzar una inferencia a nivel estatal se estableció un tamaño de muestra de 2 500 secciones. Debido a las restricciones presupuestales para poder llevar a cabo una verificación de esta magnitud, la opción elegida fue emplear la misma muestra de secciones que se visitaron en la Verificación Nacional Muestral 2003[30]. Al emplear la misma muestra de secciones que en el 2003, se aprovechó la información del recorrido cartográfico realizado dos años antes. Con esto, se tuvo un ahorro en la Encuesta de Cobertura, específicamente en la elaboración del marco muestral referente a la segunda etapa de muestreo (relación de manzanas habitadas).

[30] IFE-RFE, "Verificación Nacional Muestral 2005. Objetivos, Indicadores y Etapas de selección de la muestra. Segunda versión", 6 de abril de 2005.

1.5.8.3. Etapas de muestreo

Primera etapa de muestreo. Se empleó la muestra de secciones seleccionadas para la realización de la Verificación Nacional Muestral 2003[31]; en aquella ocasión, el método de muestreo empleado para la selección de secciones fue con probabilidad proporcional a su población estimada, con reemplazo. Dado que, el proceso de selección tenía la característica del reemplazo, se obtuvo una muestra con secciones repetidas. En el siguiente cuadro se presenta la distribución de las secciones seleccionadas en el 2003 según número de repeticiones y tipo de sección.

VNM03 Distribución de las secciones en muestra según número de repticiones y según tipo de sección

Seleccionadas	Sin repetición	Con una repetición	Con dos repeticiones
2424	2352	68	4
Seleccionadas	Urbanas	Rurales	Mixtas
2424	1427	641	356

En el 2003, en 11 secciones no fue posible obtener información porque se presentaron problemas que ponían en riesgo la seguridad del personal que las visitó o bien, se les impidió el paso. En esta ocasión, previo al operativo de campo, se excluyeron cuatro secciones de la muestra, dos por ser zonas militares y dos porque en su momento se sustituyeron. Las otras siete secciones se incluyeron en la muestra, y quedó a la consideración de los Vocales Estatales si era posible levantar la información.

[31] IFE-RFE, "Verificación Nacional Muestral 2003. Informe de resultados", 8 de mayo de 2003.

Etapas secundarias de muestreo

A. Encuesta de Cobertura

- **A.1 Selección de manzanas**[32]

  Para seleccionar manzanas fue necesario hacer un análisis del amanzanamiento en las áreas urbanas, el cual se basó en los resultados del recorrido cartográfico que se hizo en 2003 y en la comparación de la cartografía digitalizada de 2003 y 2005. Con el propósito de seleccionar manzanas con viviendas habitadas, se estimaron viviendas para cada manzana vigente en 2005, para lo cual se emplearon los siguientes insumos:

  - Número de viviendas habitadas según información del primer recorrido cartográfico para la VNM03.
  - Manzanas vigentes en 2003 según archivos digitales.
  - Manzanas vigentes en 2005 según archivos digitales.

  El procedimiento para estimar viviendas fue el siguiente:

  - Para las secciones urbanas y mixtas de la muestra, se comparó el plano de manzanas de 2003 con el de 2005. De esta comparación se obtuvieron las áreas de intersección generadas al sobreponer la cartografía de 2005 a la cartografía de 2003.
  - Se obtuvo la proporción de intersección: %Inter. = área de intersección/área 2003
  - Se estimó el número de viviendas habitadas para cada *área de intersección*; multiplicando el porcentaje de intersección por el número de

[32] IFE-RFE, "Verificación Nacional Muestral 2005. Encuesta de Cobertura. Selección de manzanas en secciones urbanas y mixtas", 29 de abril de 2005.

viviendas habitadas según datos de 2003.

Viv Hab $_{\text{por Area de intersección}} = \%\text{Inter.}^*$ VivHab_{2003}

- La cifra de viviendas habitadas para cada manzana vigente de 2005 se obtuvo sumando la estimación de viviendas de cada una de las *áreas de intersección* en las que participaba.
- Si las manzanas de nueva creación tenían al menos un empadronado se les asignaron viviendas habitadas, suponiendo que, para cada una de estas manzanas, el promedio de empadronados por vivienda es el mismo que el de la sección.

El procedimiento para seleccionar las manzanas fue sistemático con arranque aleatorio, las manzanas con viviendas habitadas de cada sección se ordenaron en forma ascendente respecto al número de viviendas habitadas.

a) En las secciones urbanas se seleccionaron cinco manzanas, si la sección tenía menos de cinco manzanas, se seleccionaron todas. Adicionalmente, se seleccionaron dos manzanas de reemplazo, en caso de que alguna de las primeras cinco no tuviera viviendas habitadas.

b) En las secciones mixtas se seleccionaron tres manzanas, si la sección tenía menos de tres manzanas se seleccionaron todas. Adicionalmente, se seleccionaron dos manzanas de reemplazo, en caso de que alguna de las primeras tres no tuviera viviendas habitadas.

De este procedimiento resultaron seleccionadas 11,093 manzanas, 8,371 para visitarse de manera obligatoria y 2,722 para visitarse sólo en caso de que

alguna de las manzanas obligadas no tuviera viviendas habitadas.

- **A.2 Selección de localidades**

  El marco de muestreo para la selección de localidades fueron todas aquellas que tienen soporte cartográfico al 2 de marzo de 2005 (11,060). A este marco se le hicieron ajustes con la intención de que todas las localidades habitadas tuvieran posibilidades de ser seleccionadas. El método de selección de localidades fue con probabilidad proporcional al padrón de la localidad (corte al 15 de abril).

  a) En las secciones mixtas se seleccionaron tres localidades, si la sección tenía menos de tres localidades, se seleccionaron todas.

  b) En las secciones rurales se seleccionaron seis localidades, si la sección tenía menos de seis localidades se seleccionaron todas.

  De este procedimiento resultaron seleccionadas 2,549 localidades.

- **A.3 Selección de viviendas.**

  En cada manzana en muestra se seleccionaron tres viviendas con muestreo aleatorio simple, si la manzana tenía tres o menos viviendas se seleccionaron todas. Este procedimiento lo realizó el entrevistador. Después de recorrer la manzana y contar el número de viviendas habitadas, éste seleccionó las viviendas empleando una tabla de números aleatorios.

  1. **En área rural.**

     En secciones con menos de 500 registros en padrón: En cada localidad se eligieron tres viviendas con muestreo aleatorio simple, si la localidad tenía tres o menos viviendas se seleccionaron todas.

  2. **En área urbana.**

     En secciones de 500 o más registros en padrón: En cada localidad se

eligieron dos manzanas con muestreo aleatorio simple y en cada manzana dos viviendas con muestreo aleatorio simple.

B. Encuesta de Actualización

El método de selección de registros fue sistemático con arranque aleatorio. En todas las secciones se seleccionaron 30 registros. Para cada sección los registros se ordenaron en forma ascendente, empleando como primer criterio la fecha de nacimiento y como segundo criterio la fecha del último trámite. De este procedimiento resultaron seleccionados 74,850 registros.

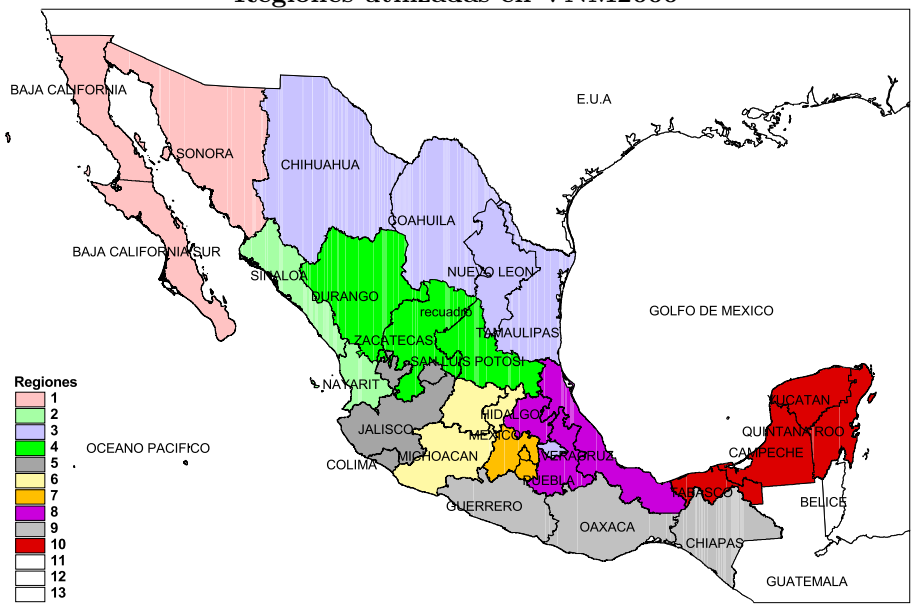
1.6. Motivación y objetivo del proyecto.

A partir de 1994, el Instituto Federal Electoral ha llevado a cabo diversos ejercicios cuyo objetivo es evaluar la calidad del padrón con base en métodos estadísticos. Los diferentes esquemas de muestreo, descritos con detalle previamente, hacen que el alcance de la inferencia no siempre sea igual, sin embargo a nivel nacional, siempre es posible hacer inferencia. Por otra parte en varios de los trabajos se ha realizado el muestreo en dos estratos: secciones urbanas y secciones no urbanas (1997, 2000, 2003, 2005).

Se ha podido observar que en varios de estos trabajos se han utilizado distintas regionalizaciones del país. Particularmente, en la VNM2000, se consideró la regionalización propuesta por CONAPO, la cual surgió como producto de indicadores sociodemográficos. Estas regiones fueron:

Región	**Estados**
1.- Noroeste I	Baja California, Baja California Sur y Sonora.
2.- Noroeste II	Sinaloa y Nayarit.
3.- Norte	Chihuahua, Coahuila, Nuevo León y Tamaulipas.
4.- Norte centro	Durango, San Luis Potosí y Zacatecas.
5.- Occidente	Aguascalientes, Colima y Jalisco.
6.- Centro	Guanajuato, Michoacán y Querétaro.
7.- Metropolitana	Distrito Federal, México y Morelos.
8.- Oriente	Hidalgo, Puebla, Tlaxcala y Veracruz.
9.- Sur	Chiapas, Guerrero y Oaxaca.
10.- Península	Campeche, Quintana Roo, Tabasco y Yucatán.

Regiones utilizadas en VNM2000



* Regiones 11, 12 y 13 corresponden a las fronteras, Golfo de México y Océano Pacífico.

Por otra parte, en la Verificación Nacional Muestral del 2002, se consideró conveniente que, previo a la selección de unidades en la muestra, se dividiera al país en cinco regiones conformadas por grupos de estados; ya que el propósito era construir modelos de regresión y por lo tanto era más factible encontrar buenos

modelos de manera regional que en el total del país. En esa ocasión se utilizaron las siguientes cinco regiones:

Región	**Estados**
I	Baja California, Baja California Sur, Coahuila, Chihuahua, Nuevo León, Sonora y Tamaulipas.
II	Durango, Nayarit, San Luis Potosí, Sinaloa y Zacatecas.
III	Aguascalientes, Colima, Guanajuato, Jalisco, Michoacán y Querétaro.
IV	Distrito Federal, Hidalgo, México, Morelos, Puebla, Tlaxcala y Veracruz.
V	Campeche, Chiapas, Guerrero, Oaxaca, Quintana Roo, Tabasco y Yucatán

Regiones utilizadas en VNM2002

* Regiones 11, 12 y 13 corresponden a las fronteras, Golfo de México y Océano Pacífico.

En este sentido, es lógico cuestionarse sobre el hecho de que si los resultados e inferencias de la VNM2000 (que se realizó con fines electorales), fueron coherentes con el resultado esperado a nivel regional. Es decir, si los valores de los princi-

pales indicadores utilizados en la VNM2000, realmente reflejaron coherencia con la realidad electoral en cada región considerada.

El objetivo general de este proyecto es proponer una regionalización de los Estados Unidos Mexicanos con las siguientes características:

- que sea útil principalmente para efectos electorales,
- con regiones NO necesariamente conexas y conformadas por entidades federativas, y
- que la realización de inferencias a nivel regional así como la estimación de indicadores, sea con mayor precisión.

Para tal efecto, utilizaré los datos que arrojaron las Verificaciones Nacionales Muestrales de 2005 y 2006, en particular, información sobre los principales indicadores electorales. Utilizaré estas fuentes, no sólo por ser recientes, sino por que tienen las siguientes ventajas:

- Mantienen el marco conceptual del estudio y
- Se tiene información a nivel entidad federativa por indicador.

Para constatar y verificar el comportamiento de los datos, emplearé herramientas de Estadística Multivariada [33], que permitan, por un lado discriminar la información en grupos homogéneos y que al mismo tiempo sean exhaustivas. En este sentido, las regiones (conformadas por entidades federativas) no necesariamente serán conexas (tal y como sucede con la regionalización de CONAPO), mas sin embargo serán homogéneas en cuanto a características electorales reflejadas en los principales indicadores.

[33] Las principales técnicas a utilizar se describen con detalle en el Capítulo 3.

Capítulo 2

Justificación

2.1. Regionalizaciones de la República Mexicana.

Las regionalizaciones que existen sobre la República Mexicana son muy distintas y variadas. La más simple y referida a tiempos remotos divide al país, propiamente a Norteamérica, en dos grandes espacios: Árido-América y Mesoamérica. Los criterios utilizados en esta regionalización fueron de tipo cultural y climático (Palerm, 1979).

Para épocas más recientes existen muchos intentos de regionalización, pero tres propuestas han sido bastante aceptadas en el medio académico, la del geógrafo mexicano Ángel Bassols (1992), la del sociólogo capitalino Luis Unikel (1976) y la del geógrafo francés Claude Bataillón (1990). Sus principales diferencias radican en una manera distinta de resolver los problemas de siempre: la articulación de los estados norteños en dirección norte-sur o en dirección este-oeste así como la definición y los límites de lo que serían la región occidente con respecto a la del centro[1]. La coincidencia radica en la caracterización de la región sur o sureste y

[1] Según Duran, J., Enero 2005.

los estados que la conforman.

Por otro lado, los estudiosos de la migración han recurrido a clasificar y analizar su información de acuerdo a criterios regionales. La mayoría utiliza o adapta regionalizaciones ya establecidas (Escobar, et.al. 1999), otros clasifican la información de acuerdo a sus propios criterios o intereses (Verduzco, 1998), unos más utilizan criterios geográficos (Lozano, 2000); mientras que Durand (1998(b); 2001) propuso una regionalización, donde articula criterios geográficos y migratorios y subdivide el territorio mexicano en cuatro grandes regiones: **histórica, fronteriza, central y sureste**.

Regiones migratorias de origen

Posteriormente Rodolfo Corona (2000, pp.183) retoma la clasificación propuesta por Durand y sólo cambia los nombres de las regiones: a la llamada Región Histórica le llamó Región Tradicional y a la que se nombró Región Fronteriza le

llamó Región Norte. Por su parte, el Consejo Nacional de Población (CONAPO) ha obtenido regionalizaciones con base en el *Índice de Marginación*, el cual es una medida que se concentra en las carencias de la población respecto al acceso a bienes y servicios básicos, captados en cuatro dimensiones: Educación, Vivienda, Ingresos y Dispersión de la población.

2.2. La regionalización de CONAPO.

2.2.1. El concepto de marginación.

La *marginación*[2] es un fenómeno estructural que se origina en la modalidad, estilo o patrón histórico de desarrollo; ésta se expresa, por un lado, en la dificultad para propagar el progreso técnico en el conjunto de la estructura productiva y en las regiones del país, y por el otro, en la exclusión de grupos sociales del proceso de desarrollo y del disfrute de sus beneficios. Los procesos que modelan la marginación conforman una precaria estructura de oportunidades sociales para los ciudadanos, sus familias y comunidades, y los expone a privaciones, riesgos y vulnerabilidades sociales que a menudo escapan al control personal, familiar y comunitario y cuya reversión requiere el concurso activo de los agentes públicos, privados y sociales.

No obstante su carácter multidimensional, algunas de las formas e implicaciones demográficas y territoriales de la marginación pueden ser aproximadas a través de medidas sintéticas. Dichas medidas analítico-descriptivas son suma-

[2] En tanto el concepto de pobreza refiere *hogares* con ingresos insuficientes para adquirir una canasta básica, el índice de marginación mide el impacto global de las privaciones que la *población* padece en educación básica, condiciones de vivienda y las derivadas de la dispersión demográfica así como la remuneración de hasta dos salarios mínimos a la población ocupada.

mente útiles para la planeación del desarrollo, dado que permiten diferenciar unidades territoriales según la intensidad de las privaciones que padece su población, así como establecer órdenes de prioridad en las políticas públicas orientadas a mejorar la calidad de vida de la población y a fortalecer la justicia distributiva en el ámbito regional.

2.2.2. El Índice de Marginación.

El Índice de Marginación (IM) es una medida-resumen que permite diferenciar entidades federativas y municipios según el impacto global de las carencias que padece la población, como resultado de la falta de acceso a la educación, la residencia en viviendas inadecuadas, la percepción de ingresos monetarios insuficientes para adquirir una canasta básica y las relacionadas con la residencia en localidades pequeñas.

Se han realizado estimaciones del IM en 1970, 1980, 1990, 2000 y 2005; los Índices de Marginación de 1980, 1990 y 2000 se obtuvieron tomando como fuente de información los resultados definitivos del X, XI y XII Censos Generales de Población y Vivienda[3] respectivamente. El Índice de Marginación 2005 se llevó a cabo utilizando dos fuentes de información, el II Conteo de Población y Vivienda 2005 y la Encuesta Nacional de Ocupación y Empleo (ENOE) 2005. Estas fuentes proporcionan la información para todas las entidades federativas y municipios del país en un mismo año de observación, asimismo tienen la ventaja de que permiten mantener el marco conceptual, las dimensiones, formas de exclusión e indicadores del Índice de Marginación estimado. En este sentido, mediante el IM es posible captar una gran parte de las manifestaciones económicas de la marginación en

[3] Correpondientes a 1980, 1990 y 2000 respectivamente.

función de las privaciones a que está expuesta la población por el hecho de residir en determinadas zonas del país. Por ejemplo, una comparación del IM por entidad federativa para 1980 y 1990[4] deja ver que los rezagos sociales en ocho indicadores disminuyeron pese a la crisis económica. La excepción fue el aumento en 13 % del porcentaje de la población que padeció privaciones por ocupar viviendas en condiciones de hacinamiento. Con ello se produjo un desplazamiento hacia los estratos intermedios, ya que la reducción en 30 % de las entidades que en 1980 se ubicaron con muy alta y alta marginación benefició al estrato medio que pasó de nueve a tres entidades.

Por último, se registro una mejoría y a la vez una persistencia de la marginación, pues si bien en Aguascalientes y Coahuila mejoraron las condiciones (de baja a muy baja) en Chiapas, Oaxaca, Veracruz, Guerrero e Hidalgo, la población siguió expuesta a vivir en condiciones de muy alta marginación.

2.2.3. El Índice de Marginación del año 1990: su construcción.

Para la estimación del Índice de Marginación del 1990, se consideró como fuente de información el XI Censo General de Población y Vivienda (correspondiente a 1990), debido a que contaba con la cobertura, grado de desagregación y actualidad de los datos necesarios para la construcción del índice a nivel estatal y municipal. En primera instancia, se identificó cada uno de los tabulados censales que contenían la información necesaria para el cálculo de nueve indicadores socioeconómicos, concentrados en cuatro dimensiones previamente mencionadas, a

[4]Los IM de 1980 y 1990 publicados por CONAPO no son directamente comparables. El ajuste y cálculo de sus varianzas explicadas a nivel de entidades federativas fue realizado en Pamplona, por Francisco y Vianey Capuzano, *Inequidad y rezago en las condiciones de salud. Diferencias regionales y estatales. SSA, México, 1994. Inédito.*

través de las cuales, se captan las carencias de la población en el acceso a bienes y servicios básicos.

- **Educación.** La escolaridad de la población constituye uno de los factores decisivos para aumentar la productividad del trabajo e incorporar la innovación tecnológica, y con ello fortalecer la competitividad de las economías. Sin duda la mayor intensidad de la marginación social, derivada de la falta de participación en el sistema educativo, se registra en la población que carece de los conocimientos que pueden adquirirse en el primer nivel de la educación básica, cuya desventaja se acentúa entre los adultos. En atención a estas consideraciones, los indicadores de educación que reflejan los rezagos más significativos, así como la proporción de la población en mayor desventaja, son:

  a) Porcentaje de la población de 15 años o más que es analfabeta.

  b) Porcentaje de población de 15 años o más sin educación primaria completa.

- **Vivienda.** La vivienda es el espacio afectivo y físico donde los cónyuges, hijos u otros parientes cercanos, estructuran y refuerzan sus vínculos familiares a lo largo de las distintas etapas de su curso de vida. Asimismo, la vivienda constituye un espacio determinante para el desarrollo de las capacidades y opciones de las familias y de cada uno de sus integrantes para llevar a cabo sus proyectos de vida. De esta manera, el alojamiento en una vivienda digna y decorosa, el cual es un derecho sancionado en el Artículo Cuarto Constitucional, favorece el proceso de integración familiar en un marco de respeto a las individualidades, evita el hacinamiento, contribuye a la creación de un clima educacional favorable para la población en edad escolar, reduce los

riesgos que afectan la salud, y facilita el acceso a los sistemas de información y entretenimiento modernos.

En cuatro de los cinco indicadores construidos para medir la intensidad de la marginación social relacionada con las condiciones de vivienda, se tomó como referente a la población no participante. Los indicadores correspondientes a esta dimensión son:

a) Porcentaje de ocupantes en viviendas particulares sin agua entubada. La falta de agua entubada propicia la utilización del vital líquido en condiciones perjudiciales para la salud debido a las formas de almacenamiento que comúnmente utilizan los residentes de este tipo de viviendas, lo que además obliga a los miembros de los hogares a invertir tiempo y esfuerzo físico en el acarreo del agua, al tiempo que dificulta el desempeño de las labores domésticas.

b) Porcentaje de ocupantes en viviendas particulares sin drenaje ni sanitario exclusivo. La falta de estos servicios en la vivienda aumenta la vulnerabilidad al incrementar el riesgo de contraer enfermedades transmisibles como las gastrointestinales y respiratorias, afectando la calidad de vida no sólo de las personas que ocupan las viviendas sin esas condiciones, sino también la de quienes comparten el hábitat, de forma que la defecación al aire libre o la carencia de sistemas para el desalojo de las aguas negras y sucias genera grandes problemas de salud pública.

c) Porcentaje de ocupantes en viviendas particulares sin disponibilidad de energa eléctrica. La carencia de electricidad excluye a la

población del disfrute de bienes culturales, del uso de los sistemas modernos de comunicación y entretenimiento, así como de la utilización de aparatos electrodomésticos. Ello también redunda en el uso de fuentes de energía alternas con altos costos ambientales y financieros.

d) Porcentaje de ocupantes en viviendas particulares con piso de tierra. Las viviendas sin ningún tipo de recubrimiento en el piso limitan las oportunidades de las personas para gozar de una vida larga y saludable, y elevan sensiblemente el riesgo de fallecer de los menores de edad por contagio de enfermedades gastrointestinales y respiratorias, principalmente donde es más difícil el acceso a los servicios de salud.

e) Porcentaje de viviendas particulares con algún nivel de hacinamiento. Conforme lo establecido por diversos organismos internacionales, se considera que en una vivienda existe hacinamiento cuando duermen en un cuarto más de dos personas; está condición compromete además la privacidad de las personas ocupantes de viviendas particulares, propiciando espacios inadecuados para el estudio y el esparcimiento, entre otras actividades esenciales para el desarrollo de las personas.

- **Ingresos por trabajo.** Las oportunidades de las personas para tener un nivel de vida digno están determinadas por una diversidad de factores. De ellos destacan: la posesión de activos, el acceso a satisfactores esenciales relacionados con el gasto social del estado (como la educación y la salud), así como las posibilidades de lograr una participación competitiva en los mercados de trabajo. El indicador correspondiente a esta dimensión incluido para la construcción del índice de marginación es:

Porcentaje de población ocupada con ingresos de hasta dos salarios mínimos. En las economías donde el mercado desempeña un papel cada vez más determinante en la asignación de los recursos escasos, el ingreso monetario determina las capacidades para adquirir bienes y servicios. Aún cuando poderosos factores extraeconómicos influyen en la determinación de los salarios, las remuneraciones guardan relación con la productividad del trabajo, sobre todo en el caso de los ingresos de los trabajadores con bajas calificaciones.

- **Dispersión de la población.** Una localidad es todo lugar ocupado con una o más viviendas, las cuales pueden estar habitadas o no; este lugar es reconocido por la ley o la costumbre. De acuerdo con sus características y con fines estadísticos, se clasifican en urbanas y rurales. El hecho de residir en localidades pequeñas, dispersas y en situación de aislamiento, no sólo hace difícil aprovechar las economías de escala[5] de los servicios básicos, de la infraestructura y el equipamiento, sino que por razones de costo-beneficio ha determinado que las acciones de la política social se concentren en la atención de quienes viven en las grandes concentraciones urbanas. Esas circunstancias *crean* una relación (inversa) entre el tamaño del asentamiento y la carencia de los servicios básicos. Con la finalidad de integrar en el índice de marginación esta dimensión de la exclusión social, se incorpora el siguiente indicador:

Porcentaje de población que vive en localidades con menos de 5,000 habitantes.

[5] La *economía de escala* se refiere al poder que tiene una empresa cuando alcanza un nivel óptimo de producción para ir produciendo más a menor coste, es decir, a medida que la producción en una empresa crece, sus costes por unidad producida se reducen.

Una vez calculados los nueve indicadores socioeconómicos que permitieron medir cada una de las formas de exclusión previamente expuestas, fue necesario construir, a partir de ellos, una medida resumen que diera cuenta de la intensidad del fenómeno. En este sentido, se buscó generar un indicador que evalúe el impacto global del rezago o déficit y que además cumpliera ciertas características que facilitarían el análisis de la expresión territorial de la marginación, tales como:

1. Reducir la dimensionalidad original y al mismo tiempo retener y reflejar al máximo posible la información referida a la dispersión de los datos en cada uno de los nueve indicadores, así como las relaciones entre ellos, y

2. Establecer un ordenamiento entre las unidades de observación (estados, municipios o localidades).

Teniendo en cuenta este propósito, se recurrió al análisis de Componentes Principales, que es una técnica estadística que transforma un conjunto de variables (o indicadores) en uno nuevo[6], donde, con un número menor de variables, se pretende reelaborar una interpretación más sencilla del fenómeno original [7].

Para el cálculo de las componentes principales se puede utilizar la matriz de covarianzas o la matriz de correlaciones. La primera se emplea cuando las variables originales tienen aproximadamente la misma varianza, de manera que el cálculo de las componentes se hace en términos de las variables originales. La segunda se emplea cuando las escalas de medición de las variables difieren o sus varianzas son notablemente distantes. Esta segundo opción es precisamente la que se siguió para obtener los Índices de Marginación. Por otra parte, por la forma de construcción de los indicadores, se convino acotarlos al intervalo [0,100], es

[6] Conjunto que recibe el nombre de *componentes principales.*

[7] Esta técnica se describe ampliamente en el Capítulo 3.

decir, cero cuando ninguno de los habitantes de alguna unidad de análisis sufre la privación que refiere el indicador y cien cuando todos los habitantes, susceptibles, padecen dicha forma de exclusión social. Con el fin de eliminar los efectos de escala entre las variables[8], éstas se estandarizaron, y se obtuvo:

$$z_{ij} = \frac{I_{ij} - \bar{I}_j}{ds_j}$$

donde[9]:

z_{ij} : es el j-ésimo indicador estandarizado ($j = 1, \ldots, 9$), de la i-ésima unidad de observación ($i = 1, \ldots, 32$ en el caso estatal ó bien hasta el número de municipios en cuestión),

I_{ij} : es el j-ésimo indicador socioeconómico, de la i-ésima unidad de análisis,

$\bar{I}_j$: es el promedio aritmético de los valores del j-ésimo indicador, y

ds_j : es la desviación estándar insesgada del j-ésimo indicador socioeconómico.

Denotando estas nuevas variables estandarizadas como $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_9$ se obtuvieron combinaciones lineales (las componentes principales) de dichas variables estandarizadas[10]:

$$\begin{aligned}
\mathbf{Y}_1 &= a_{11}\,\mathbf{Z}_1 + a_{12}\,\mathbf{Z}_2 + \ldots + a_{19}\,\mathbf{Z}_9 \\
\mathbf{Y}_2 &= a_{21}\,\mathbf{Z}_1 + a_{22}\,\mathbf{Z}_2 + \ldots + a_{29}\,\mathbf{Z}_9 \\
&\vdots \\
\mathbf{Y}_9 &= a_{91}\,\mathbf{Z}_1 + a_{92}\,\mathbf{Z}_2 + \ldots + a_{99}\,\mathbf{Z}_9
\end{aligned}$$

Adicionalmente, las variables $\mathbf{Y}_k$ deben:

1.-No estar correlacionadas, es decir, $Cov\,(\mathbf{Y}_r, \mathbf{Y}_k) = 0$ para $r \neq k$,

[8] Aunque el recorrido de las nueve variables estuvo acotado por ambos lados, fue necesario transformarlas de tal manera que aquellas con una mayor varianza no predominaran en la determinación del índice y vuelvan inoperante el análisis multivariado.

[9] Nótese que estas nuevas variables z_{ij} tienen media cero y varianza igual a uno.

[10] Los coeficientes a_{ij} reciben el nombre de *coeficientes de ponderación.*

2.- Se ordenan de tal manera que $Var(\mathbf{Y}_1) \geq Var(\mathbf{Y}_2) \geq \ldots \geq Var(\mathbf{Y}_9)$,

3.- Se eligen coeficientes de tal manera que cada vector $\mathbf{a}_k$ esté normalizado:

$$\|\mathbf{a}_k\| = \mathbf{a}_k^{'}\,\mathbf{a}_k = \sum_{i=1}^{9} a_{ik}^2 = 1$$

En las aplicaciones a nivel estatal y municipal que se derivaron de los datos del II Conteo de Población y Vivienda 2005 y de la Encuesta Nacional de Ocupación y Empleo (ENOE) 2005, se utilizó el paquete SPSS10, el cual proporciona Componentes Principales Estandarizados, con media cero y desviación estándar uno. Para ello, se reestiman los coeficientes de ponderación:

$$c_{ij} = \frac{a_{ij}}{\sqrt{\lambda_j}}$$

donde λ_j es el j-ésimo valor propio asociado a la matriz de varianzas y covarianzas, $\mathbf{V}$, de los datos estandarizados, z_{ij}.

De esta manera, los índices de marginación corresponden a la primera componente estandarizada de cada nivel de análisis, la cual como se mencionó, es una combinación lineal de las nueve variables estandarizadas, a decir:

$$\mathbf{y}_{i1} = \sum_{j=1}^{9} c_{1j} z_{ij} = c_{11} z_{i1} + c_{12} z_{i2} + \cdots + c_{19} z_{i9} = IM_i$$

donde:

y_{i1} : es el valor de la i-ésima unidad de análisis en la primera componente principal estandarizada,

c_{1j} : es el ponderador del j-ésimo indicador para determinar la primera componente principal estandarizada,

z_{ij} : es el j-ésimo indicador estandarizado de la i-ésima unidad de análisis, e

IM_i : es el valor del Índice de Marginación en la i-ésima unidad de análisis.

Cabe hacer notar que para calcular los Índices de Marginación estatales y municipales se pudo haber aplicado una metodología diferente, sin embargo, se optó por Componentes Principales debido a razones conceptuales, programáticas y teć nicas[11]. Desde una perspectiva conceptual y según se mencionó anteriormente, la marginación es un fenómeno complejo y multidimensional que tiene múltiples formas de expresión, entre las que se incluyen la insuficiencia del ingreso, la falta de acceso a los conocimientos que brinda la educación y la carencia de una vivienda apropiada, entre otras.

Desde el punto de vista programático y de instrumentación de políticas públicas, fue necesario disponer de instrumentos analíticos que permitieran sintetizar esta complejidad de orden conceptual en una medida resumen que posibilite ordenar y diferenciar a las entidades federativas y los municipios del país según la intensidad de las privaciones que afectan a la población. La técnica de componentes principales permite recuperar tanto la multidimensionalidad conceptual del fenómeno de la marginación, como la posibilidad, a través de la consideración de la primera componente, de tener un índice resumen del fenómeno para cada uno de los estados y municipios.

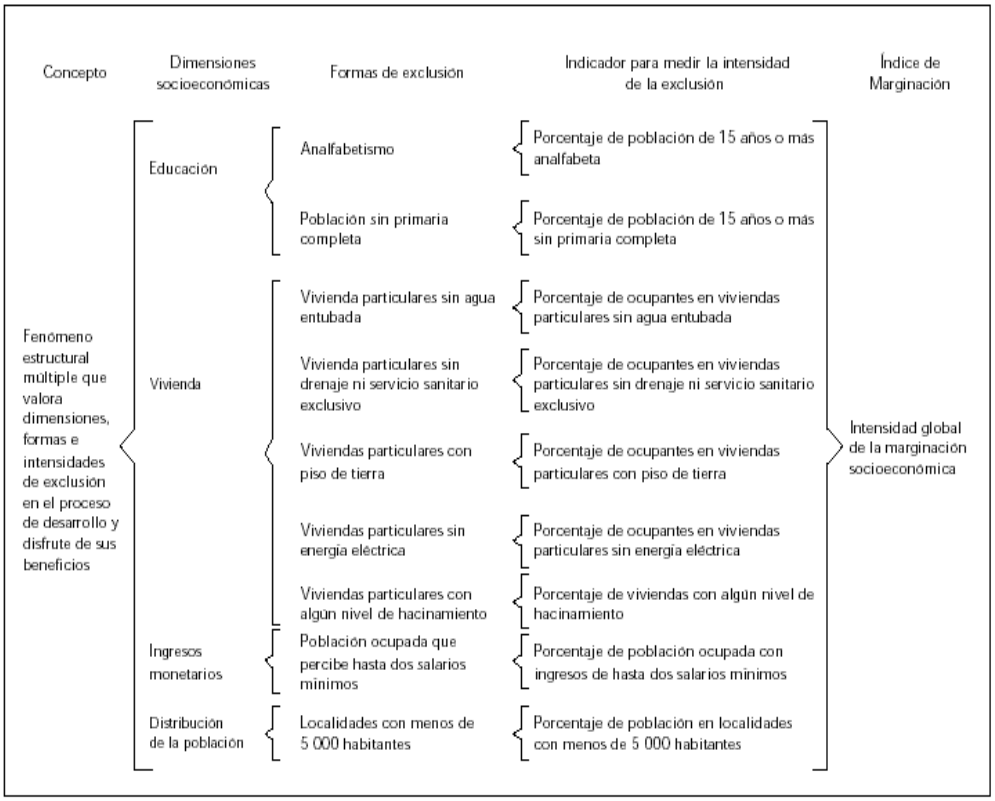
Se tuvo especial cuidado de incorporar en el análisis, exclusivamente variables cuya expresión empírica tuviera un claro referente conceptual como indicadores de la marginación social. Como resultado, se presentó correlación estadística entre cada uno de los nueve indicadores incorporados.

Finalmente, además de que la primera componente principal proporciona una

[11] Anexo C, "Metodología de estimación del índice de marginación".

medida resumen de la información de los nueve indicadores de marginación, es también un índice que recupera de la mejor manera la estructura de variación de dichos indicadores, en tanto que corresponde a la combinación sintética que explica la mayor variabilidad del conjunto de variables originales. El siguiente esquema muestra de manera conceptual la obtención del Índice de Marginación.

Esquema conceptual de la marginación



2.2.3.1. Aplicación a nivel estatal.

Debido a que se contó con información a nivel estatal, los nueve indicadores socioeconómicos que dan cuenta de las formas de exclusión social, son variables de rezago o déficit, esto es, indican el nivel relativo de privación en el que se subsumen importantes contingentes de población en cada entidad federativa.

Por ello y tal como ya se mencionó, una vez realizado el *análisis de componentes principales*, se procedió a estimar los coeficientes que ponderan cada una de las variables estandarizadas, con objeto de obtener la primera componente principal, precisamente, el Índice de Marginación (IM). Esta ordenación de los coeficientes se refleja en el porcentaje de variación de cada indicador, que es explicado por la primera componente principal. Los coeficientes obtenidos permitieron calcular el Índice de Marginación para cada entidad federativa, como una combinación lineal de los indicadores estandarizados. Este índice conllevó una fuerte ordenación de los estados, ya que estaba construido en una escala de intervalo; y esta cualidad del índice así como la aplicación de estos procedimientos estadísticos, permitieron identificar cinco estratos de marginación claramente diferenciados: **muy baja, baja, media, alta y muy alta**. Para ello, se utilizó la Técnica de Estratificación Óptima desarrollada por Dalenius y Hodges[12,13].

2.2.3.2. Aplicación a nivel municipal.

En el ámbito municipal también resultó importante dimensionar las carencias de la población, ésto con el propósito de orientar de manera eficiente los recursos públicos para atender dichas necesidades, principalmente en estos niveles político-administrativos. Cabe mencionar que en esta escala de análisis (municipal), el promedio de los nueve indicadores se magnificó debido a la mayor desigualdad intermunicipal, respecto a la desigualdad interestatal.

De igual manera como se describió para los valores estatales, se procedió a obtener la primera componente principal (el Índice de Marginación), como com-

[12] Esta técnica se describirá en el Apéndice B.

[13] *Cfr.*, Avila Jose Luis, *et.al.*,"El concepto de marignación", en CONAPO, *Indicadores socioeconómicos e índice de marginación municipal, 1990.* México, CNA-CONAPO, 1993, pp. 7-20

binación lineal de las variables municipales, ya estandarizadas, junto con los coeficientes de ponderación. La ordenación que se desprendió de los valores de los nueve coeficientes se reflejó en el porcentaje de variación de cada indicador que fue explicado por la primera componente principal. Estos coeficientes permitieron calcular el Índice de Marginación para cada municipio o delegación (en el caso del D.F.) y tal como ocurrió con las entidades federativas, la Técnica de Estratificación Óptima llevó a dividir el recorrido del Índice de Marginación municipal en cinco grupos de marginación: **muy bajo, bajo, medio, alto o muy alto.**

2.2.4. Resultados sobre el Índice de Marginación de 1990.

En relación a la construcción del IM del año de 1990, la *percepción de ingresos monetarios insuficientes* fue la variable con menos influencia en la diferenciación municipal y estatal en 1990, en tanto que las primeras dos fueron el analfabetismo y la primaria incompleta de la población adulta. Esto es así porque en ese año, el 63.22 % de la población ocupada obtenía ingresos insuficientes y se distribuía con cierta homogeneidad, hecho que puede constatarse en que el Índice de Gini[14] del indicador para los 2,403 municipios fue de 0.085[15]. Como puede verse en el siguiente mapa, el IM de 1990 por entidad federativa, permitió definir 10 grandes regiones; esta regionalización guarda notables semejanzas con las formuladas en 1970 y 1980, así como con otras realizadas a partir de criterios geoeconómicos, urbanos y flujos migratorios[16] .

[14] Hay muchas formas de definir el Índice de Gini, una de ellas es la razón entre el área de concentración y el área de máxima concentración. Por lo tanto, el Índice de Gini asumirá el valor cero si el ingreso está equidistribuido y el valor uno si está totalmente concentrado. Este concepto se describirá en el Apéndice A.

[15] *Cfr.*, Avila Jose Luis, *et.al.*,*Desigualdad regional y marginación municipal en México*, 1990, CONAPO-CNA, México, 1995

[16] Síntesis de algunas de las principales conclusiones presentadas en CONAPO. *Desigualdad regional y marginación municipal en México, 1990*, CONAPO-CNA, México, 1995.

Región 1 Noroeste I: Baja California, Baja California Sur y Sonora.

Región 2 Noroeste II: Sinaloa y Nayarit.

Región 3 Norte: Chihuahua, Coahuila, Nuevo León y Tamaulipas.

Región 4 Norte Centro: Durango, San Luis Potosí y Zacatecas.

Región 5 Occidente: Aguascalientes, Colima y Jalisco.

Región 6 Centro: Guanajuato, Michoacán y Querétaro.

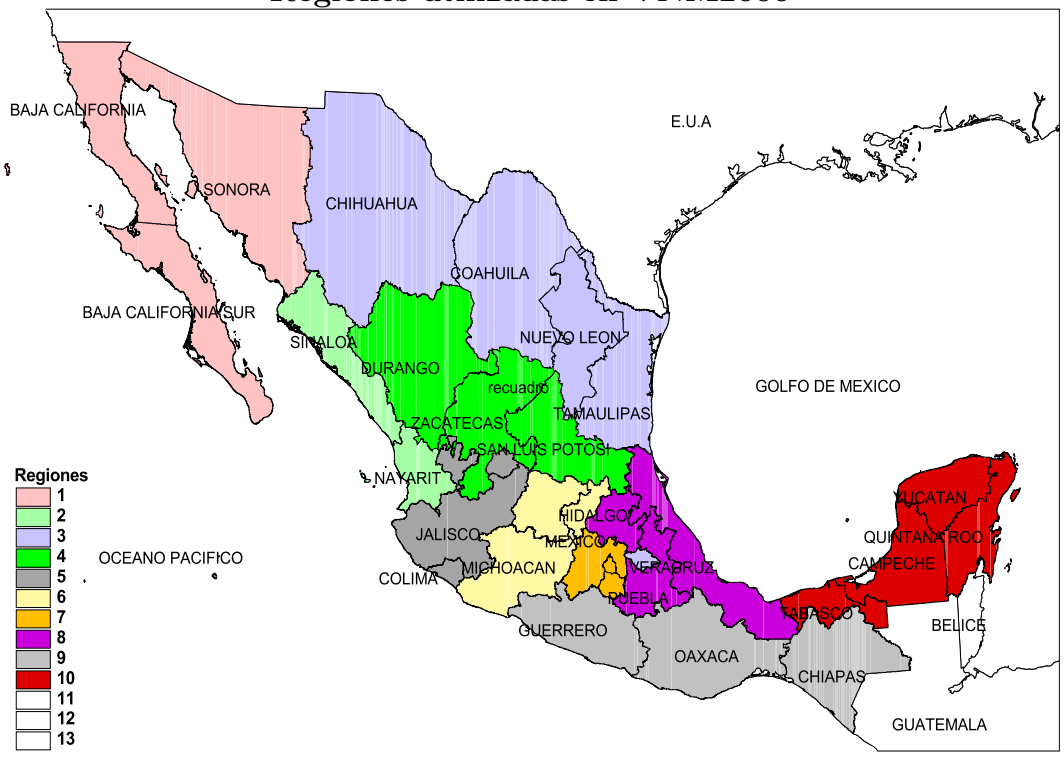
Región 7 Metropolitana: Distrito Federal, Estado de México y Morelos.

Región 8 Oriente: Hidalgo, Puebla, Tlaxcala y Veracruz.

Región 9 Sur: Chiapas, Guerrero y Oaxaca.

Región 10 Península: Campeche, Quintana Roo, Tabasco y Yucatán.

Regiones utilizadas en VNM2000



Por otra parte, el IM a nivel municipal permitió identificar las zonas con mayor polarización social, así como las exclusiones específicas que padece la población. Estratificando los municipios de cada una de las diez regiones, pueden identificarse las 17 microregiones más vulnerables del país, en las que, hasta 1990, habitaban 11.5 millones de personas que representaron el 14.18 % de la población total, e incluyeron al 93 % de los municipios identificados en 1990 como de muy alta marginación[17].

Aún con estas 17 micro - regiones caracterizadas además por una alta polaridad social interior, la diversidad de situaciones geográficas y sociales se impuso exigiendo entonces la focalización de las intervenciones públicas y privadas. En consecuencia, la marginación comprometió el nivel de vida de millones de mexicanos. Su persistencia en las micro-regiones de mayor rezago hizo factible su transmisión de padres a hijos. Por su diversidad (geográfica y cultural) y atraso productivo común, fue necesario que las intervenciones públicas y privadas consideraran la participación de los marginados en la toma de decisiones, a fin de incorporar en los programas de recuperación productiva tanto el entorno geoeconómico como las necesidades específicas de la población.

2.2.5. Resumen: El Índice de Marginación de CONAPO.

El concepto de *marginación* en su versión más abstracta intenta dar cuenta del acceso diferencial de la población al disfrute de los beneficios del desarrollo. La medición se concentra en las carencias de la población de las localidades en el

[17] Síntesis de algunas de las principales conclusiones presentadas en CONAPO. *Desigualdad regional y marginación municipal en México, 1990*, CONAPO-CNA, México, 1995.

acceso a los bienes y servicios básicos, captados en cuatro dimensiones: educación, vivienda, ingresos y dispersión de la población. Debe notarse que la marginación es un fenómeno que afecta a las localidades y no necesariamente a las personas que habitan en ella. En efecto, una localidad puede ser de muy alta marginación pero algunos de sus habitantes pueden ser alfabetos, vivir en viviendas con aguas entubadas, energía eléctrica, piso firme, bajo índice de hacinamiento y obtener ingresos suficientes como para no ser considerados pobres.

El Índice de Marginación de CONAPO es un valioso instrumento para orientar la política, pues la base de datos incluye además los nueve indicadores, variables que permiten situar en el mapa del país las localidades y entidades federativas según su grado de marginación.

Finalmente, nótese que la obtención de IM por parte de CONAPO, no se llevó a cabo con fines electorales, sino con la intención de conocer los distintos grados de marginación en México con base en indicadores de tipo socio - demográfico. Este hecho junto con la descripción proporcionada de la construcción del IM, motiva más aún la realización del proyecto, y se espera que la regionalización resultante sea más coherente para estudios con fines electorales (estadísticos, socio demográficos, etc).

Capítulo 3

Técnicas Multivariadas.

3.1. Componentes Principales.

3.1.1. Introducción.

Un problema central en el análisis de datos multivariantes es la reducción de la dimensionalidad: si es posible describir con precisión los valores de p variables con un pequeño subconjunto $r<p$ de ellas, se habrá reducido la dimensión del problema a costa de una pequeña pérdida de información. El *análisis de componentes principales* tiene este objetivo: dadas n observaciones de p variables, se analiza si es posible, representar adecuadamente esta información con un número menor de variables, construidas como combinaciones lineales de las originales. Esta técnica es debida a Hotelling (1933), aunque sus orígenes se encuentran en los ajustes ortogonales por mínimos cuadrados introducidos por K. Pearson (1901). Su utilidad es doble:

1. Permite representar óptimamente en un espacio de dimensión pequeña, observaciones de un espacio general p-dimensional. En este sentido, el *análisis de componentes principales* es el primer paso para identificar las posibles

variables no observadas que generan los datos.

2. Permite transformar las variables originales, en general correlacionadas, en nuevas variables *no correlacionadas*, facilitando así la interpretación de los datos, a costa de una pequeña pérdida de información.

Supongamos que se dispone de los valores de p variables, a decir, $\mathbf{X_1}, \mathbf{X_2}, \ldots,$ $\mathbf{X_p}$ en n elementos $I_1, I_2, \ldots, I_n$ de una población, dispuestos en una matriz $\mathbf{X}$ de dimensiones $n \times p$, donde las columnas contienen las variables y las filas los elementos o individuos; de la siguiente manera:

	$\mathbf{X_1}$	$\mathbf{X_2}$	$\cdots$	$\mathbf{X_p}$
I_1	X_{11}	X_{12}	$\cdots$	X_{1p}
I_2	X_{21}	X_{22}	$\cdots$	X_{2p}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
I_n	X_{n1}	X_{n2}	$\cdots$	X_{np}

$$\longrightarrow \mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}$$

Sin pérdida de generalidad, se puede suponer que previamente se ha restado a cada variable $\mathbf{X_i}$ su media, de tal forma que las variables de la matriz $\mathbf{X}$ tienen media cero y su matriz de varianzas y covarianzas, $\mathbf{S}$, estará dada por $\mathbf{X'X}/n$. En efecto, si denotamos $\mathbf{S} = \mathbf{X'X}/n$, entonces:

$$\begin{aligned} \mathbf{S} &= \frac{1}{n} \begin{pmatrix} X_{11} - \bar{X}_1 & X_{21} - \bar{X}_1 & \cdots & X_{n1} - \bar{X}_1 \\ X_{12} - \bar{X}_2 & X_{22} - \bar{X}_2 & \cdots & X_{n2} - \bar{X}_2 \\ \vdots & \vdots & \ddots & \vdots \\ X_{1p} - \bar{X}_p & X_{2p} - \bar{X}_p & \cdots & X_{np} - \bar{X}_p \end{pmatrix} \begin{pmatrix} X_{11} - \bar{X}_1 & X_{12} - \bar{X}_2 & \cdots & X_{1p} - \bar{X}_p \\ X_{21} - \bar{X}_1 & X_{22} - \bar{X}_2 & \cdots & X_{2p} - \bar{X}_p \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} - \bar{X}_1 & X_{n2} - \bar{X}_2 & \cdots & X_{np} - \bar{X}_p \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} \left(X_{i1} - \bar{X}_1\right)^2 & \frac{1}{n}\sum_{i=1}^{n} \left(X_{i1} - \bar{X}_1\right)\left(X_{i2} - \bar{X}_2\right) & \cdots & \frac{1}{n}\sum_{i=1}^{n} \left(X_{i1} - \bar{X}_1\right)\left(X_{ip} - \bar{X}_p\right) \\ \frac{1}{n}\sum_{i=1}^{n} \left(X_{i2} - \bar{X}_2\right)\left(X_{i1} - \bar{X}_1\right) & \frac{1}{n}\sum_{i=1}^{n} \left(X_{i2} - \bar{X}_2\right)^2 & \cdots & \frac{1}{n}\sum_{i=1}^{n} \left(X_{i2} - \bar{X}_2\right)\left(X_{ip} - \bar{X}_p\right) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n}\sum_{i=1}^{n} \left(X_{ip} - \bar{X}_p\right)\left(X_{i1} - \bar{X}_1\right) & \frac{1}{n}\sum_{i=1}^{n} \left(X_{ip} - \bar{X}_p\right)\left(X_{i2} - \bar{X}_2\right) & \cdots & \frac{1}{n}\sum_{i=1}^{n} \left(X_{ip} - \bar{X}_p\right)^2 \end{pmatrix} \end{aligned}$$

$$= \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_1, X_p) & \mathrm{Cov}(X_2, X_p) & \cdots & \mathrm{Var}(X_p) \end{pmatrix}$$

El problema que se desea resolver es encontrar un espacio de menor dimensión que represente adecuadamente los datos. Este temática puede abordarse desde tres perspectivas equivalentes:

a) Enfoque descriptivo.

Se desea encontrar un subespacio de dimensión menor que p tal que, al proyectar los puntos sobre él, éstos conserven su estructura con la menor distorsión posible. Para concretar, consideremos el caso de dos dimensiones, p=2.

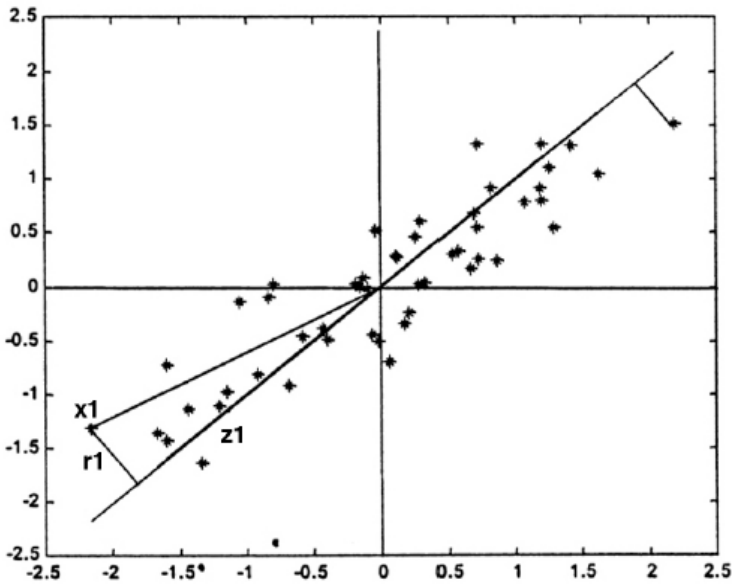


figura 1

La **figura 1** indica el diagrama de dispersión de un conjunto de puntos (que representan datos) y una recta que, intuitivamente, proporciona un buen resumen

de los datos, ya que ésta pasa "cerca" de todos los puntos. La condición de que la recta pase "cerca" de la mayoría de los puntos puede concretarse exigiendo que las distancias entre los puntos originales y sus proyecciones sobre la recta sean lo más pequeñas posibles. En consecuencia, si se considera un punto (vector que representa a un individuo) $\mathbf{I}_k = (X_{k1}, X_{k2}, \ldots, X_{kp})$, $k = 1, \ldots, n$, y una dirección $\mathbf{a}_1 = (a_{11}, \ldots, a_{1p})'$ de norma unidad, entonces la proyección del punto $\mathbf{I}_k$, sobre esta dirección es el escalar[1]:

$$z_k = \frac{\langle \mathbf{I}'_k, \mathbf{a}_1 \rangle}{\|\mathbf{a}_1\|_2^2} = \mathbf{I}_k \cdot \mathbf{a}_1 = \sum_{j=1}^{p} a_{1j}\, X_{kj} \qquad (CP.1)$$

y el vector que representa la dirección[2] será $z_k \cdot \mathbf{a}_1$. Llámese r_k, $k = 1, \ldots, n$, a las distancias entre el punto $\mathbf{I}_k$, y su proyección sobre la dirección $\mathbf{a}_1$. Este criterio (enfoque descriptivo) implica resolver el problema:

$$\text{mín} \sum_{k=1}^{n} r_k^2 = \sum_{k=1}^{n} \|\mathbf{I}'_k - z_k\, \mathbf{a}_1\|_2^2$$

Asímismo, la **figura 1** muestra que al proyectar cada punto sobre la recta se forma un triángulo rectángulo donde la hipótenusa es la distancia del punto al origen, $\|\mathbf{I}_k\|_2$, y los catetos la proyección del punto sobre la recta, z_k, y la distancia entre el punto y su proyección, r_k. Por el Teorema de Pitágoras se obtiene que:

$$\sum_{k=1}^{n} \|\mathbf{I}_k\|_2^2 = \sum_{k=1}^{n} z_k^2 + \sum_{k=1}^{n} r_k^2 \qquad \Longleftrightarrow \qquad \sum_{k=1}^{n} r_k^2 = \sum_{k=1}^{n} \|\mathbf{I}_k\|_2^2 - \sum_{k=1}^{n} z_k^2$$

Luego entonces, minimizar $\sum_{k=1}^{n} r_k^2$, la suma de las distancias de todos los puntos a la recta, es equivalente a maximizar $\sum_{k=1}^{n} z_k^2$, la suma de los cuadrados de la proyecciones. Ahora, recuérdese que las $p = 2$ variables, $\mathbf{X}_1$ y $\mathbf{X}_2$, tienen media

[1] Los cálculos se realizan en el espacio producto interno $\mathbb{R}^p$ considerando el producto interno canónico $\langle \cdot, \cdot \rangle : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, tal que $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}'\mathbf{y}$.

[2] Entendiendo que el k-ésimo individuo es el vector fila $\mathbf{I}_k = (X_{k1}, X_{k2}, \ldots, X_{kp})$.

cero, lo cual implica que los coeficientes, z_k, de las proyecciones son variables con media cero; en efecto:

$$\mathbf{E}(z_k) = \mathbf{E}\left(\sum_{j=1}^{p} a_{1j} X_{kj}\right) = \sum_{j=1}^{p} a_{1j} \cdot \mathbf{E}(X_{kj}) = 0 \qquad \forall k = 1, \ldots, n$$

pues $\mathbf{E}(X_{kj}) = 0, \forall k = 1, \ldots, n$. Por consiguiente, maximizar la suma de cuadrados de los z_k, equivale a maximizar su varianza; y de esta forma se obtiene el criterio de encontrar la dirección de proyección que maximice la varianza de los datos proyectados.

■

De manera intuitiva, se puede observar que la recta "ajustada" en **figura** 1 parece adecuada por que conserva lo más posible la variabilidad de los puntos. Sin embargo, si en lugar de buscar la dirección que pasa cerca de los puntos buscamos la dirección tal que los puntos proyectados sobre ella conserven lo más posible sus distancias relativas, llegamos al mismo criterio. En efecto, si se denota $d_{ij}^2 = \langle \mathbf{I}_i, \mathbf{I}_j \rangle$ a los cuadrados de las distancias originales entre los puntos (individuos) y $\hat{d}_{ij}^2 = (z_i - z_j)^2$ a las distancias entre los puntos proyectados sobre la recta, se desea que:

$$D = \sum_i \sum_j \left(d_{ij}^2 - \hat{d}_{ij}^2\right) = \sum_i \sum_j d_{ij}^2 - \sum_i \sum_j \hat{d}_{ij}^2$$

sea mínima. Como la suma de las distancias originales,$\sum_i \sum_j d_{ij}^2$, es fija, entonces minimizar D equivale a maximizar $\sum_i \sum_j \hat{d}_{ij}^2$, que es precisamente la distancia entre los puntos proyectados.

Proposición. *Maximizar las distancias al cuadrado entre los puntos proyectados equivale a maximizar la varianza de la variable definida por las proyecciones de los puntos.*

Prueba. Considérese el espacio $\mathbb{R}^p$ y el producto interno canónico $\langle \cdot, \cdot \rangle : \mathbb{R}^p \times \mathbb{R}^p \longrightarrow \mathbb{R}$. Sea:

$$z_k = \frac{\langle \mathbf{I}'_k, \mathbf{a}_1 \rangle}{\|\mathbf{a}_1\|_2^2} = \langle \mathbf{I}'_k, \mathbf{a}_1 \rangle$$

la proyección de la observación (individuo) $\mathbf{I}_k$ sobre la dirección $\mathbf{a}_1$, donde se supone que $\langle \mathbf{a}_1, \mathbf{a}_1 \rangle = 1$. Recuérdese que cada variable $\mathbf{X}_j$, $j = 1, \ldots, n$ tiene media cero, lo cual implica que $\sum_{k=1}^{n} X_{kj} = 0$, entonces:

$$\sum_{k=1}^{n} z_k = \sum_{k=1}^{n} \left(\sum_{j=1}^{p} a_{1j} \, X_{kj} \right) = \sum_{j=i}^{p} \left[a_{1j} \underbrace{\left(\sum_{k=1}^{n} X_{kj} \right)}_{=0} \right]$$

lo cual demuestra que cada variable z_k tiene media cero. Por otro lado, la suma de distancias al cuadrado entre los puntos proyectados esta dada por:

$$\begin{aligned} D_p &= \sum_{i=1}^{n} \sum_{h=i+1}^{n} (z_i - z_h)^2 \\ &= (n-1) \sum_{i=1}^{n} z_i^2 - 2 \sum_{i=1}^{n} \sum_{h=i+1}^{n} z_i \, z_h \\ &= n \sum_{i=1}^{n} z_i^2 - \underbrace{\left(\sum_{i=1}^{n} z_i^2 + 2 \sum_{i=1}^{n} \sum_{h=i+1}^{n} z_i \, z_h \right)}_{=B} \\ &= n \sum_{i=1}^{n} z_i^2 - B \end{aligned}$$

Ahora nótese que

$$\begin{aligned} B &= \sum_{i=1}^{n} z_i^2 + 2 \sum_{i=1}^{n} \sum_{h=i+1}^{n} z_i \, z_h \\ &= \left(z_1^2 + z_2^2 + \ldots + z_n^2 \right) + \\ &\quad 2 \left[(z_1 \, z_2 + \ldots + z_1 \, z_n) + (z_2 \, z_3 + \ldots + z_2 \, z_n) + \ldots + (z_{n-1} \, z_n) \right] \\ &= \left(z_1^2 + z_1 \, z_2 + \ldots + z_1 \, z_n \right) + \left(z_2 \, z_1 + z_2^2 + \ldots + z_2 \, z_n \right) + \\ &\quad \ldots + \left(z_n \, z_1 + z_n \, z_2 + \ldots + z_n^2 \right) \end{aligned}$$

$$
\begin{aligned}
&= z_1 (z_1 + z_2 + \ldots + z_n) + z_2 (z_1 + z_2 + \ldots + z_n) + \\
&\quad \ldots + z_n (z_1 + z_2 + \ldots + z_n) \\
&= \left(\sum_{i=1}^{n} z_i \right) \left(\sum_{i=1}^{n} z_i \right) \\
&= 0
\end{aligned}
$$

Por consiguiente, maximizar las distancias entre los puntos equivale a maximizar la expresión:

$$f(\mathbf{x}_i) = n \sum_{i=1}^{n} z_i^2$$

que coincide con el criterio de **maximizar la varianza** de la nueva variable obtenida[3].

■

b) Enfoque estadístico.

Representar puntos p - dimensionales con la mínima pérdida de información en un espacio de dimensión uno es equivalente a sustituir las p variables originales por una nueva variable z_1, que resuma óptimamente la información. Esto supone que la nueva variable debe tener, globalmente, máxima correlación con las originales o, en otros términos, debe permitir prever las variables originales con la máxima precisión. Esto no será posible si la nueva variable toma un valor semejante en todos los elementos.

A continuación, se demostrará que la condición para que sea posible prever los datos observados con la mínima pérdida de información, es utilizar la variable de

[3] Algunos autores han propuesto minimizar $\sum \sum \omega_{ij} \left(d_{ij} - \hat{d}_{ij} \right)^2$ donde ω_{ij} es una función de ponderación. El problema así planteado no tiene solución simple y debe resolverse mediante un algoritmo iterativo no lineal (Krzanowski, 1990; Cap. 2).

máxima variabilidad. De manera precisa, **se probará que los componentes principales son predictores óptimos de las variables X.**

Proposición. *La aproximación óptima de la matriz* $\mathbf{X}$ *de rango* p, *por otra matriz* $\hat{\mathbf{X}}_r$ *de dimensión* $n \times r$, *con* $r < p$, *y de rango máximo es* $\hat{\mathbf{X}}_r = \mathbf{X}\mathbf{A}_r\mathbf{A}_r'$ *donde la matriz* $\mathbf{A}_r$ *es de* $p \times r$ *y sus columnas son los vectores propios asociados a los* r *valores propios mayores de la matriz* $\mathbf{S}$, *donde* $\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i'/n$ *y* $\boldsymbol{X}_i$ *es el vector* $p \times 1$ *de observaciones en el i-ésimo elemento de la muestra (nótse que* $\boldsymbol{S}$ *es la matriz de varianzas y covarianzas).*

Prueba. El problema de aproximar la matriz $\mathbf{X}$ puede establecerse de la siguiente manera: consideremos un espacio de dimensión r definido por una base $\mathbf{U}_r$ ortonormal[4], por ende, si $\mathbf{U}_r$ es de $p \times r$ entonces $\mathbf{U}_r'\mathbf{U}_r = \mathbf{I}$. Se desea encontrar un aproximación de la matriz $\mathbf{X}$ utilizando esta base, es decir, se quiere prever cada una de las filas $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ de la matriz a través de los vectores $\mathbf{U}_r$, donde $\mathbf{X}_i$ es el vector de $p \times 1$ de observaciones en el i-ésimo individuo (o elemento) de la muestra. La predicción de la variable $\mathbf{X}_i$ será la proyección ortogonal sobre el espacio generado por estos vectores, a decir:

$$\hat{\mathbf{X}}_i \quad = \quad \mathbf{U}_r\mathbf{U}_r'\mathbf{X}_i \qquad\qquad (EE.1)$$

y se desea determinar los vectores $\mathbf{U}_r$ de tal manera que el error cuadrático de estas predicciones sea mínimo. Es claro que el error cuadrático para todos los elementos de la matriz $\mathbf{X}$ viene dado por:

$$E_{cuad} = \sum_{j=1}^{p}\sum_{i=1}^{n}\left(X_{ij} - \hat{X}_{ij}\right)^2 = \sum_{i=1}^{n}\left(\mathbf{X}_i - \hat{\mathbf{X}}_i\right)'\left(\mathbf{X}_i - \hat{\mathbf{X}}_i\right)$$

[4]Ésto puede asumirse sin pérdida de generalidad, solo recuérdese el Procedimiento de Gram-Schmidt.

el cual se quiere minimizar. Notemos que esta última igualdad podemos re-escribirla, utilizando $(EE.1)$, como:

$$\begin{aligned} E_{cuad} &= \sum_{i=1}^{n} \left\langle \mathbf{X}_i - \hat{\mathbf{X}}_i, \mathbf{X}_i - \hat{\mathbf{X}}_i \right\rangle \\ &= \sum_{i=1}^{n} \left(\mathbf{X}_i - \hat{\mathbf{X}}_i \right)' \left(\mathbf{X}_i - \hat{\mathbf{X}}_i \right) \\ &= \sum_{i=1}^{n} \left(\mathbf{X}_i' \mathbf{X}_i - 2\, \mathbf{X}_i' \hat{\mathbf{X}}_i + \hat{\mathbf{X}}_i' \hat{\mathbf{X}}_i \right) \\ &= \sum_{i=1}^{n} \left(\mathbf{X}_i' \mathbf{X}_i - 2\, \mathbf{X}_i' \mathbf{U}_r \mathbf{U}_r' \mathbf{X}_i + \mathbf{X}_i' \mathbf{U}_r \underbrace{\mathbf{U}_r' \mathbf{U}_r}_{=\mathbf{I}} \mathbf{U}_r' \mathbf{X}_i \right) \\ &= \sum_{i=1}^{n} \mathbf{X}_i' \mathbf{X}_i - 2 \sum_{i=1}^{n} \mathbf{X}_i' \mathbf{U}_r \mathbf{U}_r' \mathbf{X}_i \end{aligned}$$

De esta manera, minimizar el error cuadrático, E_{cuad}, equivale a maximizar el segundo término del miembro izquierdo de la última igualdad. De manera paralela, utilizando el hecho de que $\sum_{i=1}^{n} \mathbf{X}_i' \mathbf{U}_r \mathbf{U}_r' \mathbf{X}_i$ es un escalar y las propiedades de la traza de una matriz, se obtiene:

$$\begin{aligned} \sum_{i=1}^{n} \mathbf{X}_i' \mathbf{U}_r \mathbf{U}_r' \mathbf{X}_i &= \text{tra} \left(\sum_{i=1}^{n} \mathbf{X}_i' \mathbf{U}_r \mathbf{U}_r' \mathbf{X}_i \right) \\ &= \sum_{i=1}^{n} \text{tra} \left(\mathbf{U}_r \mathbf{U}_r' \mathbf{X}_i \mathbf{X}_i' \right) \\ &= \text{tra} \left(\mathbf{U}_r \mathbf{U}_r' \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i' \right) \\ &= n \cdot \text{tra} \left(\mathbf{U}_r \mathbf{U}_r' \mathbf{S} \right) \\ &= n \cdot \text{tra} \left(\mathbf{U}_r' \mathbf{S} \mathbf{U}_r \right) \end{aligned}$$

Esta última igualdad se debe al siguiente hecho: pre-multiplicar o post-mulitiplicar por una misma matriz, no altera el valor de la traza del producto. Entonces, de acuerdo con los cálculos anteriores se tiene que:

$$\sum_{i=1}^{n} \mathbf{X}_i' \, \mathbf{U}_r \, \mathbf{U}_r' \, \mathbf{X}_i = n \cdot \text{tra} \left(\mathbf{U}_r' \, \mathbf{S} \, \mathbf{U}_r \right)$$

Según esta expresión, minimizar el error cuadrático, E_{cuad}, implica encontrar un conjunto de vectores, $\mathbf{U}_r = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r]$, que maximicen la suma de los elementos en la diagonal de la matriz $\mathbf{U}_r' \, \mathbf{S} \, \mathbf{U}_r$, es decir, que maximicen $\sum_{j=1}^{n} \mathbf{u}_j' \mathbf{S} \mathbf{u}_j$.

Tal y como se abordará en la siguiente sección, si $r = 1$, éste es el problema que consiste en encontrar el primer componente; si $r = 2$, como el nuevo vector debe ser ortogonal al primero, se obtiene el segundo componente, y así sucesivamente.

■

c) Enfoque geométrico

El problema puede abordarse desde un punto de vista geométrico con el mismo resultado final. Si consideramos la nube de puntos de la **figura** 1, se puede observar que éstos se situán siguiendo una elipse y podemos describirlos por su proyección en la dirección del eje mayor de la elipse. Es posible demostrar que este eje es la recta que minimiza estas distancias ortogonales, con lo cual nos regresamos al problema que ya se ha resuelto. En varias dimensiones sen tendrán elipsoides y la mejor aproximación es la proporcionada por su proyección sobre el eje mayor del elipsoide. Intuitivamente, la mejor aproximación en dos dimensiones es la proyección sobre el plano de los dos ejes mayores del elipsoide y así sucesivamente. Considerar los ejes del elipsoide como nuevas variables originales supone pasar de variables correlacionadas a variables ortogonales o no correlacionadas, como se verá a continuación.

3.1.2. Cálculo de las componentes principales.

3.1.2.1. Cálculo del primer componente principal.

Al aplicar la técnica de componentes principales el objetivo es encontrar una transformación ortogonal de las variables originales a variables no correlacionadas y ordenadas de manera tal que, la primera componente (primera variable del vector resultante de la transformación) retiene la mayor parte de la variabilidad presente en todas las observaciones. En esencia, se está en busca de una transformación ortogonal, digamos $\mathbf{A}$, tal que las componentes de $\mathbf{Z} = \mathbf{XA}$ sean independientes y además $Var(\mathbf{z}_1) \geq Var(\mathbf{z}_2) \geq \ldots \geq Var(\mathbf{z}_p)$. Las columnas $\mathbf{z}_i$ de $\mathbf{Z}$ serán, precisamente, los componentes principales.

En este sentido, el *primer componente principal* se define como la combinación lineal de las variables originales que tiene varianza máxima. Los valores en este primer componente de los n individuos se representarán por el vector $\mathbf{z}_1$, y con base en lo anterior, obtenemos:

$$\mathbf{z}_1 = \mathbf{X} \cdot \mathbf{a}_1$$

tal que $Var\left(\mathbf{X} \cdot \mathbf{a}_1\right)$ sea máxima. Ya se ha visto que las variables originales tienen media cero, y por ende $\mathbf{z}_1$ tendrá media idéntica al vector 0. Su varianza será:

$$\frac{1}{n}\mathbf{z}_1' \, \mathbf{z}_1 = \frac{1}{n}\mathbf{a}_1' \, \mathbf{X}' \, \mathbf{X} \, \mathbf{a}_1 = \mathbf{a}_1' \, \mathbf{S} \, \mathbf{a}_1 \qquad (EE.2)$$

donde $\mathbf{S}$ es la matriz de varianzas y covarianzas. Claramente se observa que para maximizar la varianza, basta maximizar el módulo del vector $\mathbf{a}_1$. Por consiguiente, para que $(EE.2)$ tenga solución, se debe imponer una restricción al módulo del vector $\mathbf{a}_1$ y, sin pérdida de generalidad, se impondrá $\|\mathbf{a}_1\|_2^2 = 1$. Se introducirá esta restricción mediante el multiplicador de Lagrange:

$$M = \mathbf{a}_1 \, \mathbf{S} \, \mathbf{a}_1 - \lambda \left(\|\mathbf{a}_1\|_2^2 - 1 \right)$$

Derivando con respecto a los componentes de $\mathbf{a}_1$ e igualando a cero, se obtiene

$$\frac{\delta M}{\delta \mathbf{a}_1} = 2\,\mathbf{S}\,\mathbf{a}_1 - 2\,\lambda\,\mathbf{a}_1 = 0 \qquad \Longleftrightarrow \qquad \mathbf{S}\,\mathbf{a}_1 = \lambda\,\mathbf{a}_1$$

lo cual implica que $\mathbf{a}_1$ es un vector propio de la matriz $\mathbf{S}$, y λ su correspondiente valor propio. Ahora, multiplicando la última igualdad por $\mathbf{a}_1'$ (por la izquierda) se tiene:

$$\mathbf{a}_1'\,\mathbf{S}\,\mathbf{a}_1 = \left(\underbrace{\frac{1}{n}\mathbf{z}_1'\,\mathbf{z}_1}_{\mathrm{Var}(\mathbf{z}_1)} = \right) \lambda\,\mathbf{a}_1'\,\mathbf{a}_1 = \lambda\,\|\mathbf{a}_1\|_2^2 = \lambda$$

es decir λ, el correspondiente valor propio de $\mathbf{a}_1$, es igual a la varianza de $\mathbf{z}_1$; y dado que esta es la cantidad que se pretende maximizar, entonces λ será el mayor valor propio de la matriz $\mathbf{S}$. Su vector asociado, $\mathbf{a}_1$, define los coeficientes de cada variable en el primer componente principal.

3.1.2.2. Cálculo del segundo componente principal.

A partir de este componente, el objetivo es obtener el mejor plano de proyección de las variables $\mathbf{X}$, lo cual se calculará estableciendo como función objetivo el hecho de que la suma de las varianzas de $\mathbf{z}_1 = \mathbf{X}\cdot\mathbf{a}_1$ y $\mathbf{z}_2 = \mathbf{X}\cdot\mathbf{a}_2$, sea máxima, donde $\mathbf{a}_1$ y $\mathbf{a}_2$ son los vectores ortogonales (i.e. $\mathbf{a}_1'\,\mathbf{a}_2 = 0$) que definen el plano. En este sentido, la función objetivo será:

$$\phi\,(\mathbf{a}_1, \mathbf{a}_2) = \mathbf{a}_1'\,\mathbf{S}\,\mathbf{a}_1 + \mathbf{a}_2'\,\mathbf{S}\,\mathbf{a}_2 - \lambda_1\left(\|\mathbf{a}_1\|_2^2 - 1\right) - \lambda_2\left(\|\mathbf{a}_2\|_2^2 - 1\right) \qquad (EE.3)$$

que incorpora las restricciones de que las direcciones deben tener módulo unitario; y de manera análoga al cálculo del primer componente, si se deriva e iguala a cero para maximizar, se obtiene:

$$\begin{aligned} \frac{\delta \phi}{\delta \mathbf{a}_1} &= 2\,\mathbf{S}\,\mathbf{a}_1 - 2\,\lambda_1\,\mathbf{a}_1 = 0 \\ \frac{\delta \phi}{\delta \mathbf{a}_2} &= 2\,\mathbf{S}\,\mathbf{a}_2 - 2\,\lambda_2\,\mathbf{a}_2 = 0 \end{aligned}$$

La solución de este sistema es:

$$\begin{aligned} \mathbf{S} \cdot \mathbf{a}_1 &= \lambda_1 \mathbf{a}_1 \\ \mathbf{S} \cdot \mathbf{a}_2 &= \lambda_2 \mathbf{a}_2 \end{aligned}$$

lo cual indica que $\mathbf{a}_1$ y $\mathbf{a}_2$ deben ser vectores propios de $\mathbf{S}$ y λ_1 y λ_2 los valores propios asocidados respectivamente. Procediendo de manera análoga que en el cálculo del primer componente, se tiene:

$$\begin{aligned} \mathbf{a}_1' \, \mathbf{S} \, \mathbf{a}_1 &= \lambda_1 \, \|\mathbf{a}_1\| \\ \mathbf{a}_2' \, \mathbf{S} \, \mathbf{a}_2 &= \lambda_2 \, \|\mathbf{a}_2\| \end{aligned}$$

Tomando los vectores propios de norma la unidad y sustituyendo en $(EE.3)$, se obtiene que, en el máximo, la función objetivo toma el valor $\phi = \lambda_1 + \lambda_2$. Es claro que λ_1 y λ_2 deben ser los dos valores propios mayores de la matriz $\mathbf{S}$ y, $\mathbf{a}_1$ y $\mathbf{a}_2$, sus correspondientes vectores propios. En este caso, $\mathbf{a}_2$ define los coeficientes de cada variable en el segundo componente principal.

De manera paralela, nótese que la covarianza entre $\mathbf{a}_1$ y $\mathbf{a}_2$, dada por $\mathbf{a}_1 \, \mathbf{S} \, \mathbf{a}_2$, es cero ya que $\mathbf{a}_1' \cdot \mathbf{a}_2 = 0$, y por ende las variables $\mathbf{z}_1$ y $\mathbf{z}_2$ estarán **no correlacionadas.**

3.1.2.3. Generalización.

Se puede demostrar de manera análoga que el espacio de dimensión $r < p$ que mejor representa a los puntos viene definido por los vectores propios asociados a los r mayores valores propios de la matriz $\mathbf{S}$. Estas direcciones de denominan *direcciones principales* de los datos, y a las nuevas variables[5] por ellas definidas *componentes principales*[6]. En general, la matriz $\mathbf{X}$ (y por lo tanto $\mathbf{S}$) tiene ran-

[5] Se les ha llamdado $\mathbf{Z}_i$.

[6] Precisamente, las combinaciones lineales de las variables originales que se buscaban.

go p, existiendo entonces tantas componentes principales como variables que se obtendrán calculando los valores propios o raíces características, $\lambda_1, \ldots, \lambda_p$ de la matriz de varianzas y covarianzas $\mathbf{S}$, mediante:

$$|\mathbf{S} - \lambda_i \mathbf{I}| = 0$$

y sus vectores asociados son:

$$(\mathbf{S} - \lambda_i \mathbf{I})\, \mathbf{a}_i = 0$$

Al ser $\mathbf{S}$ una matriz simétrica, los valores propios λ_i's son reales y por el hecho de ser positiva definida, los λ_i son positivos. Además, si λ_j y λ_h son dos valores propios distintos, el hecho de que $\mathbf{S}$ sea simétrica implica que sus vectores propios asociados sean ortogonales. En efecto, supóngase que $\mathbf{a}_j$ y $\mathbf{a}_h$ son los vectores propios asociados a λ_j y λ_h respectivamente, entonces:

$$\mathbf{S} \cdot \mathbf{a}_j = \lambda_j \, \mathbf{a}_j \qquad y \qquad \mathbf{S} \cdot \mathbf{a}_h = \lambda_h \, \mathbf{a}_h$$

Premultipicando por $\mathbf{a}_h^{'}$ la primera igualdad, se obtiene:

$$\begin{aligned}
& \lambda_j \left(\mathbf{a}_h^t \cdot \mathbf{a}_j\right) = \mathbf{a}_h^t \cdot (\mathbf{S} \cdot \mathbf{a}_j) \\
\Longleftrightarrow \quad & \lambda_j \left(\mathbf{a}_h^t \cdot \mathbf{a}_j\right) = \left(\mathbf{a}_h^t \cdot \mathbf{S}\right) \cdot \mathbf{a}_j \\
\Longleftrightarrow \quad & \lambda_j \left(\mathbf{a}_h^t \cdot \mathbf{a}_j\right) = \left(\mathbf{a}_h^t \cdot \mathbf{S}^t\right) \cdot \mathbf{a}_j \\
\Longleftrightarrow \quad & \lambda_j \left(\mathbf{a}_h^t \cdot \mathbf{a}_j\right) = (\mathbf{S} \cdot \mathbf{a}_h)^t \cdot \mathbf{a}_j \\
\Longleftrightarrow \quad & \lambda_j \left(\mathbf{a}_h^t \cdot \mathbf{a}_j\right) = \left(\lambda_h \, \mathbf{a}_h^t\right) \cdot \mathbf{a}_j \\
\Longleftrightarrow \quad & (\lambda_j - \lambda_h)\, \mathbf{a}_h^t \cdot \mathbf{a}_j = 0
\end{aligned}$$

pero $\lambda_h \neq \lambda_j$, luego entonces $\mathbf{a}_h^t \cdot \mathbf{a}_j == 0$, con lo cual se conluye que $\mathbf{a}_h$ y $\mathbf{a}_j$ son ortogonales. En particular, si se asume que la matriz $\mathbf{S}$ es semidefinida positiva de rango $r < p$, y si $p - r$ variables fuesen combinación lineal de las demás, entonces habría solamente r raíces características positivas mientras que el resto serían

ceros, lo cual significaría que algunas variables son combinaciones lineales exactas de las otras, y por lo tanto pueden ser eliminadas antes de iniciar el análisis.

Ahora, sea $\mathbf{Z}$ la matriz cuyas columnas son los vectores que conforman las p componentes en los n individuos; estas nuevas variables están relacionadas con las originales, $\mathbf{X}$, mediante $\mathbf{Z} = \mathbf{X}\,\mathbf{A}$ donde $\mathbf{A}'\,\mathbf{A} = \mathbf{I}$. En suma, calcular los componentes principales equivale a encontrar una transformación ortogonal (en este caso la matriz $\mathbf{A}$) a las variables $\mathbf{X}$ (que juegan el papel de ejes originales) para obtener las nuevas variables $\mathbf{Z}$, no correlacionadas entre sí. Esta operación puede interpretarse geométricamente como elegir nuevos ejes coordenados, que coincidan con los "ejes naturales" de los datos; lo cual se resume en el siguiente:

Teorema. *Sea* $\boldsymbol{X}$ *la matriz en donde se dispone de los valores de* p *variables medidas en* n *individuos tal que* $\boldsymbol{E}(\boldsymbol{X}) = \boldsymbol{0}$ *y* $\boldsymbol{E}(\boldsymbol{X}'\boldsymbol{X}) = \Sigma$. *Entonces, existe una transformación ortogonal:*

$$\mathbf{U} = \mathbf{X} \cdot \beta$$

de tal forma, que la matriz de varianzas y covarianzas de $\boldsymbol{U}$ *es* $\boldsymbol{E}(\boldsymbol{U}'\boldsymbol{U}) = \Lambda$, *donde:*

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}$$

y $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ *son los valores propios de* Σ. *La* r*-ésima columna de* β, *a decir* β_r, *satisface* $(\Sigma - \lambda_r \boldsymbol{I})\,\beta_r = \boldsymbol{0}$. *La* r*-ésima columna de* $\boldsymbol{U}$, $\boldsymbol{U}_r = \boldsymbol{X} \cdot \beta_r$, *es la combinación lineal no correlacionada de* $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_{r-1}$ *con máxima varianza.*

■

La matriz $\mathbf{U}$ es precisamente la matriz que contiene las nuevas variables denominadas *componentes principales* de $\mathbf{X}$. Debe notarse que $\mathbf{U}_1 = \mathbf{X} \cdot \beta_1$ es la combinación lineal normalizada (componente principal) con máxima varianza, en efecto, si $\mathbf{U}^\star = \sum_i c_i \, \mathbf{U}_i$, donde $\sum c_i^2 = 1$ [7], entonces:

$$Var\left(\mathbf{U}^\star\right) = \sum_i c_i^2 \, \lambda_i = \lambda_1 + \sum_{i=2}^{p} c_i^2 \, \left(\lambda_i - \lambda_1\right)$$

y dado que $c_1^2 = 1 - \sum_{i=2}^{p} c_i^2$, claramente Var $(\mathbf{U}^\star)$ es máxima cuando $c_i^2 = 0$ para $i = 2, \ldots, p$. Similarmente, $\mathbf{U}_2$ es la combinación lineal normalizada *no* correlacionada con $\mathbf{U}_1$ de máxima varianza[8]; por ende, las propiedades de máxima varianza para $\mathbf{U}_3, \ldots \mathbf{U}_p$ en el mismo sentido se verifican.

3.1.3. Cálculo de las varianzas.

Teorema. *Una transformación ortogonal* $\boldsymbol{V} = \mathbf{X}\,\mathbf{C}$ *de una matriz* $\boldsymbol{X}$ *de* $n \times p$ *cuyas columnas son variables aleatorias, no modifica el valor de la varianza total ni la suma de las varianzas de las componentes.*

Prueba. Sin pérdida de generalidad, podemos asumir que:

$$\mathbf{E}\left(\mathbf{X}\right) = \mathbf{0} \qquad y \qquad \mathbf{E}\left(\mathbf{X}'\,\mathbf{X}\right) = \Sigma$$

Entonces:

$$\mathbf{E}\left(\mathbf{V}\right) = \mathbf{E}\left(\mathbf{X}\right) \cdot \mathbf{C} = \mathbf{0} \qquad \text{y}$$

$$\mathbf{E}\left(\mathbf{V}'\,\mathbf{V}\right) = \mathbf{E}\left(\left(\mathbf{XC}\right)'\,\mathbf{XC}\right) = \mathbf{C}' \cdot \mathbf{E}\left(\mathbf{X}'\,\mathbf{X}\right) \cdot \mathbf{C} = \mathbf{C}'\,\Sigma\,\mathbf{C}$$

Nótese que el determinante de la varianza de $\mathbf{V}$ es:

$$|\mathbf{C}'\,\Sigma\,\mathbf{C}| = |\mathbf{C}'||\Sigma||\mathbf{C}| = |\Sigma||\mathbf{C}'\,\mathbf{C}| = |\Sigma|$$

[7] En este sentido, $\mathbf{U}^\star$ es también una combinación lineal normalizada de $\mathbf{X}$

[8] $\mathbf{U}^\star = \sum_i c_i \cdot \mathbf{U}_i$ no correlacionada con $\mathbf{U}_1$ implica $c_1 = 0$

que es precisamente, el determinante de la varianza de $\mathbf{X}$. Por otro lado, dada una variable aleatoria unidimensional, $\mathbf{Y}$, se satisface que:

$$\mathrm{Var}\left(\mathbf{Y}\right) = E\left(\mathbf{Y}^2\right) - \left[E\left(\mathbf{Y}\right)\right]^2$$

Consideremos la matriz $\mathbf{V}$ y sus columnas, $\mathbf{V_i}$'s, entonces:

$$\mathrm{Var}\left(\mathbf{V}_i\right) \;=\; \mathbf{E}\left(\mathbf{V}_i'\,\mathbf{V}_i\right) - \mathbf{E}\left(\mathbf{V}_i\right)'\,\mathbf{E}\left(\mathbf{V}_i\right)$$

pero $\mathbf{E}(\mathbf{V})$=$\mathbf{0}$, por ende, la suma de varianzas de las columnas de $\mathbf{V}$ es:

$$\begin{aligned}\sum_i \mathrm{Var}\left(\mathbf{V}_i\right) &= \sum_i \mathbf{E}\left(\mathbf{V}_i'\,\mathbf{V}_i\right)\\ &= \mathrm{tra}\left(\mathbf{C}'\,\Sigma\,\mathbf{C}\right)\\ &= \mathrm{tra}\left(\Sigma\,\mathbf{C}'\,\mathbf{C}\right)\\ &= \mathrm{tra}\left(\Sigma\right) \qquad\qquad (CV.1)\end{aligned}$$

Ahora, considérese la descomposición espectral de Σ, a decir, $\Sigma = \Lambda'\,D\,\Lambda$, donde D=$\mathbf{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r)$, con $r \leq p$, de lo cual se tiene que:

$$\begin{aligned}\mathrm{tra}\left(\Sigma\right) &= \mathrm{tra}\left(\Lambda'\,D\,\Lambda\right)\\ &= \mathrm{tra}\left(D\,\Lambda'\,\Lambda\right)\\ &= \mathrm{tra}\left(D\right)\\ &= \sum_i \lambda_i \qquad\qquad (CV.2)\end{aligned}$$

y acuerdo con los cálculos de las secciones previas se obtiene:

$$\sum_i \lambda_i = \sum_i \mathbf{E}\left(\mathbf{X}_i'\,\mathbf{X}_i\right) = \sum_i \mathrm{Var}\left(\mathbf{X}_i\right)$$

y junto con $(CV.1)$ y $(CV.2)$ se sigue que:

$$\begin{aligned}\sum_i \mathrm{Var}\left(\mathbf{V}_i\right) &= \sum_i \mathbf{E}\left(\mathbf{V}_i'\,\mathbf{V}_i\right)\\ &= \mathrm{tra}\left(\Sigma\right)\end{aligned}$$

$$
\begin{aligned}
&= \sum_i \lambda_i \\
&= \sum_i \mathbf{E}\left(\mathbf{X}_i' \mathbf{X}_i\right) \\
&= \sum_i \mathrm{Var}\left(\mathbf{X}_i\right)
\end{aligned}
$$

■

3.2. Análisis de Conglomerados.

3.2.1. Introducción.

El *análisis de comglomerados* tiene por objeto agrupar elementos en grupos homogéneos en función de las similitudes o similaridades entre ellos. Normalmente, con esta técnica se agrupan las observaciones, pero también puede aplicarse para agrupar variables. El análisis de conglomerados estudia tres tipos de problemas:

1. *Partición de datos.* Se dispone de datos generalmente heterogéneos y se desea dividirlos en un número de grupos prefijado, de manera que:
   a) cada elemento pertenezca a uno, y sólo uno, de los grupos;
   b) todo elemento quede clasificado;
   c) cada grupo sea internamente homogéneo.
2. *Construcción de jerarquías.* Se desea proporcionar, a los elementos de un conjunto una estructura de forma jerárquica por su similitud. Una clasificación jerárquica implica que los datos se ordenan en niveles, de manera que los niveles superiores contienen a los inferiores. Estríctamente, estos métodos no definen grupos, sino la estructura de asociación en cadena que pueda existir entre los elementos.
3. *Clasificación de variables.* En problemas o experimentos donde se tiene un número considerable de variables, es de gran importancia realizar un es-

tudio exploratorio inicial para dividir las variables en grupos. Este estudio funciona como orientación para plantear los modelos formales para reducir la dimensión.

Cabe señalar que las técnicas para el desarrollo de los análisis de conglomerados siempre van a funcionar, independientemente de que existan o no los grupos.

3.2.2. Método clásico de partición: El Algoritmo de las k - medias.

Supóngase que se cuenta con una muestra de n elementos en los cuales se han medido p variables; el objetivo es dividir esta muestra en un número de grupos prefijado, k. El algoritmo de las k- medias requiere las cuatro etapas siguientes:

1. Seleccionar k puntos en el espacio p - dimensional como centros de los grupos iniciales, lo cual puede hacerse:
   a) tomando como centros los k puntos más alejados entre sí;
   b) construyendo grupos iniciales con información a *priori* y calculando sus centros, o bien seleccionando los centros a *priori*, y
   c) seleccionándalos de manera aleatoria.
2. Calcular las distancias euclideanas de cada elemento a los centros de los k grupos, y asignar cada elemento al grupo de cuyo centro esté más próximo. La asignación se realiza secuencialmente y al introducir un nuevo elemento en un grupo se recalculan las coordenadas del nuevo centro del grupo.
3. Definir un criterio de optimalidad y comprobar si reasignando alguno de los elementos mejora el criterio.
4. Si no es posible mejorar el proceso de optimalidad, entonces terminar el proceso.

El criterio de optimalidad que se utiliza en el algoritmo de las k - medias es **minimizar** la suma de cuadrados dentro de los grupos (SCG_w) para todas las p

variables, dada por:

$$SCG_w = \sum_{g=1}^{k} \sum_{j=1}^{p} \sum_{i=1}^{n_k} (x_{ijg} - \bar{x}_{jg})^2 \qquad (AC.1)$$

donde x_{ijg} es el valor de la variable j en el i-ésimo individuo del grupo g y $\bar{x}_{.jg}$ la media de esta variable en el grupo. Este criterio es equivalente a minimizar la suma ponderada de las varianzas de las variables en los grupos, ya que puede escribirse como:

$$\text{mín} \sum_{g=1}^{k} \sum_{j=1}^{p} n_g \, s_{jg}^2$$

donde n_g es el número de elementos en el grupo g y s_{jg}^2 es la varianza de la variable j en dicho grupo. Las varianzas de las variables en los grupos constituyen una medida de la heterogeneidad de la clasificación y al minimizarlas se obtienen grupos más homogéneos. Un criterio alternativo de homogeneidad sería minimizar las distancias al cuadrado entre los puntos y sus centros de grupo. Si se miden las distancias con la norma euclideana, este criterio se escribe:

$$\text{mín} \sum_{g=1}^{k} \sum_{i=1}^{n_g} \left(\mathbf{X}_{ig} - \bar{\mathbf{X}}_g\right)' \left(\mathbf{X}_{ig} - \bar{\mathbf{X}}_g\right) = \text{mín} \sum_{g=1}^{k} \sum_{i=1}^{n_g} d^2\left(i, g\right)$$

donde $d^2\left(i, g\right)$ es el cuadrado de la distancia euclideana entre el elemento i del grupo g y la media del grupo. Para verificar que estos criterios son equivalentes, recuérdese que un escalar es igual a su traza, por lo cual, es posible escribir este criterio como:

$$\begin{aligned} \text{mín} \sum_{g=1}^{k} \sum_{i=1}^{n_g} d^2\left(i, g\right) &= \text{mín} \sum_{g=1}^{k} \sum_{i=1}^{n_g} \text{tra}\left[d^2\left(i, g\right)\right] \\ &= \text{mín} \left\{ \text{tra} \left[\sum_{g=1}^{k} \sum_{i=1}^{n_g} \left(\mathbf{X}_{ig} - \bar{\mathbf{X}}_g\right)' \left(\mathbf{X}_{ig} - \bar{\mathbf{X}}_g\right) \right] \right\} \end{aligned}$$

y llamando $\mathbf{W}$ a la matriz de suma de cuadrados dentro de cada grupo,

$$\mathbf{W} = \sum_{g=1}^{k} \sum_{i=1}^{n_g} \left(\mathbf{x}_{ig} - \bar{\mathbf{x}}_g\right)' \left(\mathbf{x}_{ig} - \bar{\mathbf{x}}_g\right)$$

se tiene que:

$$\text{mín}\left[\text{tra}\left(\mathbf{W}\right)\right] = \text{mín}\, SCG_w$$

De esta manera, ambos criterios coinciden. Éste se denomina *criterio de la traza*, y fue propuesto por Ward (1963). Ahora bien, nótese que la minimización de SCG_w requeriría calcular $(AC.1)$ para todas las posibles particiones, labor claramente imposible, salvo para valores de n pequeños. El algoritmo de k-medias busca la partición óptima con la restricción de que en cada iteración sólo se permite mover un elemento de un grupo a otro y funciona como sigue:

1. Partir de una asignación inicial,
2. Comprobar si moviendo algún elemento se reduce la tra $(\mathbf{W})$,
3. Si es posible reducir tra $(\mathbf{W})$ moviendo un elemento, entonces mover dicho elemento, recalcular las medias de los dos grupos afectados por el cambio y volver a (2). Si no es posible reducir la tra $(\mathbf{W})$, terminar.

En consecuencia, el resultado del algoritmo puede depender de la asignación inicial y del orden de los elementos. El criterio de la traza tiene dos propiedades importantes:

1. *La primera es que no es invariante ante cambios de escala.* Cuando las variables vayan en unidades distintas, conviene estandarizar para evitar que el resultado del algoritmo de k - medias dependa de cambios irrelevantes en la escala de medida. Cuando vayan en las mismas unidades suele ser mejor no estandarizar, ya que una varianza mucho mayor que el resto puede deberse precisamente a que existen dos grupos de observaciones en esa variable, lo cual se puede ocultar al estandarizar.
2. La segunda propiedad es que minimizar la distancia euclideana, debido a las propiedades topológicas de ésta, *produce grupos aproximadamente esféricos.*

Por otro lado, este criterio supone variables cuantitativas y, aunque puede aplicarse si existe un pequeño número de variables binarias, cuando existen muchas de ellas que son atributos, es mejor utilizar los métodos jerárquicos que se describirán más adelante.

3.2.3. Número de grupos.

En la aplicación habitual del algoritmo de k-medias es necesario fijar el número de grupos, k. Es claro que este número no puede estimarse vía un criterio de homogeneidad, pues la forma de conseguir grupos muy homogéneos y minimizar la SCG_w es hacer tantos grupos como observaciones, con lo que $SCG_w == 0$.

Se han propuesto distintos métodos para seleccionar el número de grupos; un procedimiento que se ha utilizado con gran frecuencia, es utilizar la siguiente prueba de reducción de variabilidad:

$$F = \frac{SCG\,(k) - SCG\,(k+1)}{SCG\,(k+1)\,/(n-k-1)}$$

Ésta compara la SCG_w de k grupos con la de $k+1$, calculando la reducción de variabilidad al aumentar un grupo adicional, es decir, compara la varianza promedio con la disminución de variabilidad al aumentar un grupo. El valor obtenido se compara con una F con p y $p\,(n-k-1)$ grados de libetad[9]. Una regla empírica que da resultados razonables, sugerida por Hartigan (1975), es introducir un grupo más si este cociente es mayor a 10.

[9] Esta regla aún no es del todo justificada por que los datos no tienen por qué verificar las hipótesis necesarias para aplicar la distribución F.

3.2.4. Métodos Jerárquicos.

3.2.4.1 Distancias y similaridades.

Los métodos jerárquicos parten de una matriz de distancias o similaridades entre los elementos de la muestra y construyen una jerarquía basada en estas distancias. Una larga lista de medidas de **similaridad** han sido propuestas y pueden ser categorizadas matemáticamente en distintos modos[10]. Si $\mathbb{P}$ es una población de objetos, se puede definir a similaridad como una función C, que mapea $\mathbb{P} \times \mathbb{P}$ en $\mathbb{R}$ y satisface los siguientes axiomas:

1. $0 \leq C(r, s) \leq 1$ para todo r, s en $\mathbb{P}$,
2. $C(r, r) = 1$, para todo r en $\mathbb{P}$,
3. $C(r, s) = C(s, r)$, para todo r, s en $\mathbb{P}$.

Gower [1971a] propuso la siguiente medida de similaridad:

$$c_{rs} = \left(\sum_{j=1}^{d} c_{rsj} \right) \Big/ \sum_{j=1}^{d} w_{rsj}$$

donde c_{rsj} es una medida de similaridad entre los objetos r y s en la variable j. Aquí, w_{rsj} es igual a uno excepto cuando la comparación no es posible, como sucede con observaciones pérdidas o valores negativos para variables dicotómicas, en cuyo caso $c_{rsj} = w_{rsj} = 0$.

Si todas las variables son continuas, la distancia más utilizada es la distancia euclideana entre las variables estandarizadas univariadamente. Para decidir si estandarizar o no las variables, antes del análisis conviene tener en cuenta tanto los comentarios hechos como el objetivo del estudio. Si no se estandarizan, la distancia euclideana dependerá principalmente de las variables con valores más grandes y el resultado del análisis puede cambiar completamente al modificar su

[10] Por ejemplo usando árboles en Hartigan[1967], o bien Duran y Odell [1974, cap. 4].

escala de medida. En cambio, si se estandarizan, se esta dando *a priori* un peso semejante a las variables con independencia de su variablidad original, lo que puede no ser siempre adecuado.

Pero cuando en la muestra existen variables continuas y atributos, el problema es más complicado. Supóngase que la variable $\mathbf{X}_1$ es binaria; entonces la distancia euclideana entre dos elementos de la muestra en función de esta variable es $(\mathbf{X}_{i1} - \mathbf{X}_{h1})^2$, que tomará el valor cero si y solo si $\mathbf{X}_{i1} = \mathbf{X}_{h1}$, es decir, cuando el atributo está, o no está, en ambos elementos; y uno, si el atributo está en un elemento y no en otro. Sin embargo, la distancia entre dos elementos en una variable continua estandarizada, $(\mathbf{X}_{i1} - \mathbf{X}_{h1})^2 / s_1^2$, puede ser mucho mayor que uno, en este sentido, las variables continuas van en general, a pesar mucho más que las variables binarias. Esto puede ser aceptable en muchos casos, pero cuando por la naturaleza del problema no sea deseable, una solución es trabajar con similaridades.

Si se obtienen las similaridades entre dos elementos para cada variable, se pueden combinar en un coeficiente de similaridad global entre los dos elementos. El coeficiente propuesto por Gower es:

$$s_{ih} = \frac{\sum_{j=1}^{p} \omega_{jih}\, s_{jih}}{\sum_{j=1}^{p} \omega_{jih}}$$

donde ω_{jih} es una variable que es igual a uno si la comparación de estos dos elementos mediante la variable j tiene sentido, y será cero si no se desea incluir esa variable en la comparación. Por ejemplo, si se define la variable $\mathbf{X}_1$ como:

$$\mathbf{X}_1 := \text{La persona ha pedido un crédito}$$

que toma los valores $\mathbf{X}_1 = 1$ si ha pedido un credito y $\mathbf{X}_1 = 0$ si no, y sea $\mathbf{X}_2$ la variable que indica si lo ha devuelto o no. En este caso, al comparar la pareja de individuos (i, j), si uno cualquiera de los dos tiene un valor cero en $\mathbf{X}_1$, se asignará a la variable ω_{jih} el valor cero[11].

Las similaridades entre elementos en función de las variables cualitativas pueden construirse individualmente o por bloques. La similaridad entre dos elementos por una variable binaria será uno, si ambos tienen el atributo, y cero en caso contrario[12]. Si se asume que todos los atributos tienen el mismo peso, se puede construir una medida de similaridad entre dos elementos, A y B, contando el número de atributos presentes:

- en ambos;
- en A y no en B;
- en B y no en A;
- en ninguno de los dos elementos.

Estas cuatro cantidades forman una *tabla de asociación entre elementos*, y servirán para construir medidas de similitud o similaridad entre los dos elementos comparados. Para calcular un coeficiente de similitud entre dos individuos a partir de su tabla de asociación, se utilizan principalmente los dos criterios siguientes:

1. *Proporción de coincidencias.*

   Se calcula como el número total de coincidencias sobre el número de atributos totales n_a:

$$s_{ij} = \frac{a + d}{n_a} \qquad (AC.2)$$

[11] Si una persona no ha pedido un crédito, tiene $\mathbf{X}_1 = 0$, y no tiene sentido preouparse por $\mathbf{X}_2$.

[12] Alternativamente, se pueden agrupar las variables binarias en grupos homogéneos y tratarlas conjuntamente.

2. *Proporción de apariciones.*

   Cuando la ausencia de un atributo no es relevante, se pueden excluir las ausencias y calcular solo la proporción de veces donde el atributo aparece en ambos elementos. El coeficiente se define por:

$$s_{ij} = \frac{a}{a+b+c}$$

Aunque las dos propuestas previas son las más utilizadas, puede haber situaciones donde sean recomendables otras medidas. Se puede dar peso doble a las coincidencias, con lo que resulta:

$$s_{ij} = \frac{2(a+d)}{2(a+d)+b+c} \qquad (AC.3)$$

o bien sólo tener en cuenta las coincidencias y tomar $s_{ij} = a/(b+c)$. Finalmente, los coeficientes de similitud o similaridad para una variable contínua se construyen mediante:

$$s_{ijh} = 1 - \frac{|\mathbf{x}_{ij} - \mathbf{x}_{hj}|}{rango\,(\mathbf{X}_j)}$$

y de esta manera, el coeficiente resultante estará siempre entre cero y uno.

Una vez obtenida la similaridad global entre los elementos, se pueden transformar los coeficientes en distancias. Lo más simple es definir la distancia mediante $d_{ij} = 1 - s_{ij}$, pero esta relación puede no verificar la desigualdad triangular; sin embargo, si se calculan las similaridades mediante $(AC.2)$ y $(AC.3)$, la matriz de distancias es definida positiva, mas aún, si se define la distancia por[13] $d_{ij} = \sqrt{2\,(1 - s_{ij})}$, entonces sí se verifica la propiedad triangular. Cabe mencionar que también se pueden calcular las distancias directamente o bien, utilizar alguna medida de disimilaridad.

[13] Esta conversión se abordará y deducirá en la sección 3.3.3.1.

Una medida de **disimilaridad** para medir la cercanía entre dos puntos $\mathbf{x}$, $\mathbf{y}$ es una función Δ, que mapea $\mathbb{R}^p \times \mathbb{R}^p$ en $\mathbb{R}$ y satisface los siguientes axiomas:

(1) $\Delta(\mathbf{x}, \mathbf{y}) \geq 0$, para todo $\mathbf{x}, \mathbf{y}$ en $\mathbb{R}^p$.
(2) $\Delta(\mathbf{x}, \mathbf{y}) \equiv 0$ si y solo sí $\mathbf{x} = \mathbf{y}$.
(3) $\Delta(\mathbf{x}, \mathbf{y}) = \Delta(\mathbf{y}, \mathbf{x})$, para todo $\mathbf{x}, \mathbf{y}$ en $\mathbb{R}^p$
(4) $\Delta(\mathbf{x}, \mathbf{y}) \leq \Delta(\mathbf{x}, \mathbf{z}) + \Delta(\mathbf{y}, \mathbf{z})$, para cualesquiera $\mathbf{x}, \mathbf{y}, \mathbf{z}$ en $\mathbb{R}^p$

Los axiomas (1) y (2) implican que Δ es positiva definida y el axioma (3) implica que Δ es simétrica. Ahora, utilizando la norma norma L_p, tenemos la familia de métricas Minkowski:

$$\Delta_p(\mathbf{x}, \mathbf{y}) = \left\{ \sum_{j=1}^{p} |\mathbf{x}_j - \mathbf{y}_j|^p \right\}^{1/p}$$

Las más comunes son la métrica L_1, la métrica L_2 o euclideana, y la norma supremo ($p = \infty$):

$$\Delta_\infty(\mathbf{x}, \mathbf{y}) = \sup_{1 \leq j \leq p} |\mathbf{x}_j - \mathbf{y}_j|$$

La métrica L_1 es sencilla de evaluar y fue usada por Carmichael y Sneath (1969); la norma del supremo involucra un sensibilidad a cambios de escala en alguna de las variables. Y dadas las observaciones $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, con la norma euclideana puede definirse:

$$\begin{aligned} d_{rs} &= \Delta_2(\mathbf{x}_r, \mathbf{x}_s) \\ &= \left\{ \sum_{j=1}^{p} |x_{rj} - x_{sj}| \right\}^{1/2} \end{aligned}$$

como la similaridad (distancia) entre $\mathbf{x}_r$ y $\mathbf{x}_s$. Sin embargo, un cambio en la escala puede también producir efectos sustanciales en el rango de distancias. Para tratar de superar no solamente el problema del escalamiento, sino también los efectos

de correlación entre las variables, se ha propuesto la distancia de Mahalanobis:

$$\Delta\left(\mathbf{x}_r, \mathbf{x}_s\right) = \left\{\left(\mathbf{x}_r - \mathbf{x}_s\right)' \mathbf{S}^{-1} \left(\mathbf{x}_r - \mathbf{x}_s\right)\right\}^{1/2}$$

donde $\mathbf{S} = \sum_i \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)' / (n-1)$. Esta medida es invariante ante transformaciones de la forma $\mathbf{y}_i = \mathbf{A} \cdot \mathbf{x}_i + \mathbf{b}$[14]. Distintas versiones escaladas de la métrica L_1 han sido también utilizadas. Gower (1971a) escaló cada variable por su rango, lo cual indujo la métrica:

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

donde R_j es el rango de la variable j. La medida Bray - Curtis (1957), comúnmente utilizada en ecología, tiene la forma[15]:

$$\begin{aligned} d_{rs} &= \left(\sum_j |x_{rj} - x_{sj}|\right) \Big/ \left(\sum_j x_{rj} + \sum_j x_{sj}\right) \\ &= 1 - \frac{2\sum_j \text{mín}\left(x_{rj}, x_{sj}\right)}{\sum_j x_{rs} + \sum_j x_{sj}} \end{aligned}$$

Otra medida, que constituye un métrica para variables positivas, es la métrica Canberra:

$$d_{rs} = \frac{1}{d} \sum_j \left\{\frac{|x_{rj} - x_{sj}|}{x_{rj} + x_{sj}}\right\}$$

introducida por Lance y Williams (1966, 1976) y que es robusta ante el sesgo producido por valores atípicos (outliers).

Una medida de disimilaridad d_{rs} entre objetos r y s no requiere ser una métrica, en efecto, es posible tener $d_{rs} = 0$, donde los objetos r y s son diferentes. Sibson (1972) argumenta que la relación de orden es más importante que los

[14] Hartigan (1975: p. 63) comenta que la invarianza ante transformaciones lineales generales no es tan clara como la invarianza ante cambios en las unidades de medida de cada variable.

[15] Esta medida no es una métrica, Seber(1984, p. 392) ejercicio 7.3.

valores numéricos mismos, y una función monótona creciente de un coeficiente de similaridad mantiene dicha relación. Por esta razón, la desigualdad triangular no es un requerimiento importante, y mas aún, una transformación monótona de una métrica no necesariamente satisface la desigualdad triangular.

3.2.4.2. Algoritmos jerárquicos.

Dada una matriz de distancias y similaridades, el objetivo es clasificar los elementos en una jerarquía. Los algoritmos existentes funcionan de manera tal que los elementos son sucesivamente asignados a los grupos, donde dicha asignación es irrevocable. Los algoritmos son de dos tipos:

1. *De aglomeración.* Parten de los elementos individuales y los van agregando en grupos.

2. *De división.* Parten del conjunto de elementos y lo van dividiendo sucesivamente hasta llegar a los elementos individuales.

3.2.4.3. Métodos aglomerativos.

Los algoritmos aglomerativos siguen una misma estructura y sólo se diferencian en la forma de calcular las distancias entre grupos. Dicha estructura es:

1. Comenzar con tantas clases como elementos, dígase n. Las distancias entre las clases son las distancias entre los elementos originales.

2. Seleccionar los dos elementos más próximos en la matriz de distancias y formar con ellos una clase.

3. Sustituir los dos elementos utilizados en el paso 2. para definir la clase por un nuevo elemento que represente la clase construida. Las distancias entre

este nuevo elemento y los anteriores se calculan con alguno de los criterios que se describen a continuación.

4. Regresar al paso 2. y repetir pasos 2. y 3. hasta que se tengan todos los elementos agrupados en una única clase.

Criterios para definir distancias entre grupos.

Supóngase que se tiene un grupo A con n_a elementos y un grupo B con n_b elementos y que ambos se fusionan para crear un nuevo grupo que se denotará (AB) con $n_a + n_b$ elementos. La distancia del nuevo grupo, (AB), a otro grupo C con n_c elementos, se calcula en general con alguna de las siguientes reglas:

1. *Encadenamiento simple.* La distancia entre los dos nuevos grupos es la menor de las distancias entre grupos antes de la fusión. Es decir:

$$d\left(C; AB\right) = \text{mín}\left(d_{CA}, d_{CB}\right)$$

Una forma simple de calcular este mínimo es utilizar la expresión básica:

$$\text{mín}\left(d_{CA}, d_{CB}\right) = 1/2\left(d_{CA} + d_{CB} - |d_{CA} - d_{CB}|\right)$$

Dado que este criterio solo depende del orden de las distancias, será invariante ante transformaciones monótonas, es decir, se obtendrá la misma jerarquía aunque las distancias sean numéricamente distintas.

2. *Encadenamiento completo.* La distancia entre los dos nuevos grupos es la máxima de las distancias entre grupos antes de la fusión. Es decir:

$$d\left(C; AB\right) = \text{máx}\left(d_{CA}, d_{CB}\right)$$

De manera análoga al caso anterior, este máximo puede calcularse de manera simple utilizando la expresión:

$$\text{máx}\,(d_{CA}, d_{CB}) = 1/2\,(d_{CA} + d_{CB} + |d_{CA} - d_{CB}|)$$

Este criterio será también invariante ante transformaciones monótonas de las distancias, pues depende del orden de las distancias.

3. *Media de grupos*. La distancia entre los dos nuevos grupos es la media ponderada entre las distancias antes de la fusión. A decir:

$$d\,(C; AB) = \frac{n_a}{n_a + n_b} d_{CA} + \frac{n_b}{n_a + n_b} d_{CB}$$

Dado que se ponderan los valores de las distancias, este criterio no es invariante ante transformaciones monótonas.

4. *Método del centroide*. Se aplica generalmente sólo con variables continuas y el método consiste en igualar la distancia entre dos grupos a la distancia euclideana entre sus centros, donde se toman como centros los vectores de medias de las observaciones que pertenecen al grupo. Al unirse dos grupos se pueden calcular las nuevas distancias entre ellos sin utilizar los elementos originales.

El método de Ward.

Otro proceso para construir el agrupamiento jerárquico fue propuesto por Ward y Wishart [16]. En este método se parte de los elementos y se define una

[16] Ward argumentó que los conglomerados debían construirse de tal manera que, al fundirse dos elementos, la pérdida de información resultante de la fusión fuera mínima. En ese contexto, la cantidad de información se cuantifica como la suma de las distancias al cuadrado de cada elemento respecto al centroide del conglomerado al que pertenece.

medida global de la heterogeneidad de una agrupación de observaciones, a decir:

$$\mathbf{W} = \sum_{g=1}^{k} \sum_{i=1}^{n_g} \left(\mathbf{X}_{ig} - \bar{\mathbf{X}}_g\right) \left(\mathbf{X}_{ig} - \bar{\mathbf{X}}_g\right)'$$

donde $\bar{\mathbf{X}}_g$ es la media del grupo g. Este criterio inicia suponiendo que cada dato forma un grupo, $g = n$, y, por ende, $\mathbf{W} \equiv 0$. El siguiente paso consiste en unir los elementos que produzcan el mínimo incremento en $\mathbf{W}$[17]. En la siguiente etapa se tienen $n - 1$ grupos de los cuales, $n - 2$ conformados por un elemento y uno conformado por dos elementos. Se decide de nuevo unir dos grupos de manera tal que $\mathbf{W}$ crezca la menos posible, con lo que se pasa a $n - 2$ grupos; y así sucesivamente hasta tener un grupo único. Los valores de $\mathbf{W}$ van indicando el creciemiento al formar los grupos y pueden utilizarse para decidir cuántos grupos naturales contienen los datos.

El dendrograma.

El dendrograma o árbol jerárquico, es una representación gráfica en forma de árbol del resultado del proceso de agrupamiento. Los criterios propuestos para definir distancias tienen la propiedad de que si se consideran tres grupos, A, B y C, se verifica que:

$$d(A, C) \leq \text{máx}\{d(A, B), d(B, C)\}$$

y una medida de distancia que tiene esta propiedad se denomina *ultramétrica*. Esta propiedad es más fuerte que la propiedad triangular, ya que una ultramétrica es siempre una distancia. En efecto, si $d(A, C)$ es menor o igual que el máximo entre $d(A, B)$ y $d(B, C)$, forzosamente será menor o igual que la suma $d(A, B) + d(B, C)$. El dendrograma es la representación gráfica de una ultramétrica, y se

[17] esto implica tomar los más próximos con la distancia euclideana.

construye como sigue:

1. En la parte inferior del gŕafico se disponen los n elementos iniciales.
2. Las uniones entre los elementos se indican por tres líneas rectas: dos dirigidas a los elementos que se unen, y que son perpendiculares al eje de los elementos, y una paralela a este eje, que sitúa al nivel en que se unen.
3. El proceso se repite hasta que todos los elementos estén conectados por líneas rectas.

Si se corta el dendrograma a un nivel de distancia dado, se obtiene una clasificación del número de grupos existentes a ese nivel y los elementos que los forman; además, los dendrogramas son fáciles de interpretar y son útiles cuando los elementos tienen claramente una estructura jerárquica. Sin embargo, pueden conducir a conclusiones incorrectas principalmente por el hecho de que el dendrograma correspondiente a un conglomerado jerárquico no es único, ya que depende fuertemente de la liga utilizada así como de la estructura propia de los elementos.

3.3. Escalamiento Multidimensional.

3.3.1. Introducción.

El origen del escalamiento multidimensional se remonta a los trabajos de cuantificación de Quetelet en las Ciencias Sociales pero su nacimiento está unido a los estudios de psicología experimental, en los años cincuenta, para descubrir la similaridad entre estímulos aplicados a distintos individuos. Los métodos existentes se dividen en **métricos**[18], cuando la matriz inicial es propiamente de distancias,

[18] Los métodos métricos, también llamados coordenadas principales, utlizan las diferencias entre similitudes, mientras que los no métricos parten de que si el individuo A es más similar a B que a C, entonces A está más cerca de B que de C.

y ***no* métricos**, cuando dicha matriz es de similaridades (concepto definido y referido en secciones previas).

Las técnicas de escalamiento multidimensional son una generalización de la técnica de componentes principales, cuando en lugar de disponer de una matriz de observaciones por variables, como en componentes principales, se dispone de una matriz, $\mathbf{D}$, cuadrada $n \times n$ de distancias o disimilaridades entre los n elementos de un conjunto. El objetivo es representar esta matriz mediante un conjunto de variables ortogonales $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_p$, que se denominan *coordenadas principales*, donde $p < n$, de manera que las distancias euclideanas entre las coordenadas de los elementos respecto a estas variables sean iguales (o la más próximas posibles) a las distancias o similaridades de la matriz original. Es decir, a partir de la matriz $\mathbf{D}$, se pretende obtener una matriz $\mathbf{X}$, de dimensiones $n \times p$, que pueda interpretarse como la matriz en donde la distancia euclideana entre los elementos reproduzca, aproximadamente, la matriz de distancias inicial $\mathbf{D}$.

En general, no es posible encontrar p variables que reproduzcan exactamente las distancias iniciales, pero sí es frecuente encontrar variables que reproduzcan aproximadamente las distancias iniciales. El escalamiento multidimensional comparte con los componentes principales el objetivo de describir e interpretar los datos. Si existen muchos elementos, la matriz de similaridades tendrá muchas entradas, entonces la representación por unas pocas variables de los elementos nos permitirá entender sus características: qué elementos tienen propiedades similares, si hay elementos atípicos, etc.

3.3.2. Escalados métricos: coordenadas principales.

El análisis de componentes principales se presentó como una técnica de reducción lineal donde el objetivo principal es reemplazar un conjunto de p vectores $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p \in \mathbb{R}^n$ por un conjunto de k vectores $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_k \in \mathbb{R}^n$, con $k < p$, que son combinaciones lineales de las primeras.

Matemáticamente, la técnica de escalamiento multidimensional consiste en calcular las distancias δ_{rs} entre los puntos (unidades de estudio) $\delta_{rs} = \|\mathbf{X}_r - \mathbf{X}_s\|$ y encontrar un conjunto de k vectores, $\mathbf{Y}_i$, con distancias entre los puntos $d_{rs} = \|\mathbf{Y}_r - \mathbf{Y}_s\|$ tales que $\delta_{rs} \approx d_{rs}$ para todo $r \neq s$. En muchas aplicaciones, los vectores de distancias $\mathbf{X}_i$ no son conocidos, y δ_{rs} es simplemente una medida de *proximidad* entre los objetos r y s. Esta medida de *proximidad* no es necesaria-mente la distancia euclideana, incluso ni siquiera una medida de la distancia propiamente.

Por otro lado, también es común encontrarse con el problema de que esta *proximidad* es una distancia, pero medida con error, por lo cual, el objetivo es ahora reconstruir la configuración original de las distancias entre puntos a lo más, de manera aproximada. Recuérdese que las *proximidades* se describen usualmente como *similaridades* o *disimilaridades*, según sea la información que se tenga de las unidades de estudio.

Considérese una *disimilaridad* δ_{rs}; la matriz $\mathbf{D} = [(\delta_{rs})]$ es llamada *matriz de disimilaridad*. Se dice que $\mathbf{D}$ es euclideana si existe un configuración p-dimensional $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p$, para alguna p, tal que $d_{rs} = \delta_{rs}$. El siguiente teorema (Gower [1966], Mardia et al. [1979]) proporciona condiciones para que $\mathbf{D}$ sea euclideana.

Teorema. *Sea* $\boldsymbol{A} = [(a_{rs})]$, *donde* $a_{rs} = -\frac{1}{2}\delta_{rs}^2$. *Defínase* $b_{rs} = a_{rs} - \hat{a}_{r\cdot} - \hat{a}_{\cdot s} + \hat{a}_{\cdot\cdot}$; *en términos matriciales:*

$$\mathbf{B} = [(b_{rs})] = \left(\mathbf{I}_n - n^{-1}\,\mathbf{1}_n\,\mathbf{1}_n'\right)\mathbf{A}\left(\mathbf{I}_n - n^{-1}\,\mathbf{1}_n\,\mathbf{1}_n'\right)$$

Entonces $\boldsymbol{D}$ *es Euclideana si y solo si* $\boldsymbol{B}$ *es semipositiva definida.*

■

La demostración se presenta en Seber (1984), Capítulo 5. El **Teorema** anterior proporciona un método para construir una configuración $\{\mathbf{y}_i\}_{i=1}^k$ conocida comúnmente como método "clásico" de escalamiento multidimensional. La idea básica fue introducida por Richardson [1938], y popularizado por Togerson [1952, 1958], quien introdujo el término de *escalamiento multidimensional*. Gower[1966, 1967a] clarificó más introduciendo el nombre de *análisis de coordenadas principales* y demostró que dicha técnica guarda especial relación con el análisis de componentes principales.

Es claro que la solución clásica[19] no es única; basta aprovechar el hecho de que un cambio de origen y una rotación o reflexión no modifican las distancias entre los individuos (puntos). En efecto, si $\mathbf{L}$ es una matriz ortogonal de $p \times p$, entonces:

$$\begin{aligned}
\|\mathbf{L}\,(\mathbf{y}_r - \mathbf{y}_s)\,\|_2^2 &= (\mathbf{L}\,(\mathbf{y}_r - \mathbf{y}_s))'\,(\mathbf{L}\,(\mathbf{y}_r - \mathbf{y}_s)) \\
&= (\mathbf{y}_r - \mathbf{y}_s)'\,\mathbf{L}'\mathbf{L}\,(\mathbf{y}_r - \mathbf{y}_s) \\
&= (\mathbf{y}_r - \mathbf{y}_s)'\,(\mathbf{y}_r - \mathbf{y}_s) \\
&= \|\mathbf{y}_r - \mathbf{y}_s\|_2^2
\end{aligned}$$

de tal forma que $\mathbf{L}\,\mathbf{y}_i$ donde $i = 1, \ldots, n$, proporciona otra solución.

[19] Cuando $\mathbf{D}$ sea Euclideana, se denomina *solución clásica.*

En ocasiones, la matriz de disimilaridad no es Euclideana, y por ende no se garantiza que todos los eigenvalores de $\mathbf{B}$ sean reales y positivos. Por consiguiente el **Teorema** anterior no será válido; sin embargo, si los primeros k eigenvalores son relativamente grandes y positivos, y los restantes están cercanos a cero, entonces las filas de $\mathbf{Y}_k = \left(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(k)}\right)$ proporcionarán una configuración *razonable* en dimensión k.

El procedimiento clásico también puede ser aplicado a similaridades c_{rs}. La conversión es simple utilizando la transformación[20]:

$$\delta_{rs} = \sqrt{2\left(1 - c_{rs}\right)}$$

y tomando $a_{rs} = c_{rs}$ en el **Teorema** anterior. Ahora, si $\mathbf{A} \geq 0$, $\mathbf{B} \geq 0$, y $\mathbf{D}$ es Euclideana, se sigue de éste que[21]:

$$\begin{aligned} \|\mathbf{y}_r - \mathbf{y}_s\|^2 &= a_{rr} + a_{ss} - 2a_{rs} \\ &= 2 - 2c_{rs} \\ &= \delta_{rs}^2 \end{aligned}$$

y por lo tanto $\mathbf{y}_i$ proporciona de nuevo una solución. El proceso de construcción del escalamiento clásico se describe a continuación.

3.3.3. Construcción de las coordenadas principales.

Dada una matriz $\mathbf{X}$, $n \times p$, de individuos por variables, la nueva variable:

$$\tilde{\mathbf{X}} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{X}$$

de dimensiones $n \times p$, tiene media cero. A partir de esta matriz $\tilde{\mathbf{X}}$, se pueden construir dos tipos de matrices cuadradas y semidefinidas positivas: la matriz de

[20] Esta conversión se abordará y deducirá en la sección 3.3.3.1.

[21] La expresión "≥ 0" indica que ambas matrices son **semidefinidas positivas.**

varianzas y covarianzas $\mathbf{S}$, definida por $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}/n$ de $p\times p$ y la matriz de productos cruzados, $\mathbf{Q} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}'$ de dimensiones $n \times n$, que puede interpretarse como una matriz de similitud entre los n elementos. Para verificar esta afirmación, sea $\mathbf{y} \neq \mathbf{0}_{p\times 1}$, entonces:

$$\mathbf{y}'\left(\frac{1}{n}\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)\mathbf{y} = \frac{1}{n}\left(\tilde{\mathbf{X}}\mathbf{y}\right)'\underbrace{\left(\tilde{\mathbf{X}}\mathbf{y}\right)}_{\mathbf{z}_{n\times 1}} = \frac{1}{n}\mathbf{z}'\mathbf{z} = \frac{1}{n}\sum_{i=1}^{n} z_i^2 \geq 0$$

Por lo tanto, $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}/n$ es semidefinida positiva. De manera análoga se obtiene el resultado para $\mathbf{Q}$. Ahora, los términos de la matriz $\mathbf{Q}$, q_{ij}, contienen el producto escalar por pares de elementos, es decir:

$$q_{ij} = \sum_{s=1}^{p} x_{is}\, x_{js} = \mathbf{x}_i'\,\mathbf{x}_j$$

donde $\mathbf{x}_i$ es la i-ésima fila de $\tilde{\mathbf{X}}$. Por la expresión escalar, $q_{ij} = |\mathbf{x}_i||\mathbf{x}_j|\cos\theta_{ij}$, si los dos elementos son muy similares, entonces $\cos\theta_{ij} \simeq 1$ y q_{ij} será grande. Por el contrario, mientras más ortogonales sean los elementos, entonces $\cos\theta_{ij} \simeq 0$ y q_{ij} será pequeño; en este sentido, se puede interpretar la matriz $\tilde{\mathbf{X}}\tilde{\mathbf{X}}'$ como la matriz de similitud entre elementos. Las distancias entre las observaciones se deducen inmediatamente de esta matriz de similitud. Ahora, la distancia euclideana al cuadrado entre dos elementos es:

$$\begin{aligned} d_{ij}^2 &= \|\mathbf{x}_i - \mathbf{x}_j\|^2 \\ &= \sum_{s=1}^{p}(x_{is} - x_{js})^2 \\ &= \sum_{s=1}^{p} x_{is}^2 + \sum_{s=1}^{p} x_{js}^2 - 2\sum_{s=1}^{p} x_{is}\, x_{js} \qquad (EM.0) \end{aligned}$$

que en términos de la matriz $\mathbf{Q}$, puede escribirse como:

$$d_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij} \qquad (EM.1)$$

Por lo tanto, dada una matriz $\tilde{\mathbf{X}}$ se puede construir una matriz de similaridad $\mathbf{Q} = \tilde{\mathbf{X}}\,\tilde{\mathbf{X}}'$ y, a partir de ella, la matriz de distancias al cuadrado $\mathbf{D}$ entre elementos con ayuda de (EM.1). Si se denota $diag(\mathbf{Q})$ al vector que contiene los términos diagonales de la matriz $\mathbf{Q}$, y $\mathbf{1}$ al vector de unos, la matriz $\mathbf{D}$ viene dada por:

$$\mathbf{D} = diag(\mathbf{Q}) \cdot \mathbf{1}' + \mathbf{1} \cdot diag(\mathbf{Q})' - 2\,\mathbf{Q}$$

El problema que se tratará es precisamente, el inverso: reconstruir la matriz $\tilde{\mathbf{X}}$ a partir de una matriz de distancias al cuadrado $\mathbf{D}$, con elemenetos d_{ij}^2. Para ello, primeramente se obtendrá la matriz $\mathbf{Q}$ dada la matriz $\mathbf{D}$.

Nótese que no hay pérdida de generalidad en suponer que las variables tienen media cero. Esto es consecuencia de que las distancias entre dos puntos d_{ij}^2, no varían si se expresan las variables en desviaciones respecto la media, pues:

$$d_{ij}^2 = \sum_{s=1}^{p} (x_{is} - x_{js})^2 = \sum_{s=1}^{p} \left[(x_{is} - \bar{\mathbf{x}}_s) - (x_{js} - \bar{\mathbf{x}}_s)\right]^2$$

Además, dado que se supone que la única información existente son las distancias entre elementos, se condicionará a que la matriz $\tilde{\mathbf{X}}$ por encontrar, esté formada por variables con media cero, es decir, $\tilde{\mathbf{X}}' \cdot \mathbf{1} = \mathbf{0}$. En consecuencia, $\mathbf{Q} \cdot \mathbf{1} = \mathbf{0}$, lo cual indica que la suma de todos los elementos de una fila[22] de la matriz de similitudes, $\mathbf{Q}$, debe ser cero. Para imponer estas restricciones, de $(EM.1)$, sumando por filas, se tiene:

$$\begin{aligned}
\sum_{i=1}^{n} d_{ij}^2 &= \sum_{i=1}^{n} q_{ii} + \sum_{i=1}^{n} q_{jj} - 2\underbrace{\sum_{i=1}^{n} q_{ij}}_{0} \\
&= \sum_{i=1}^{n} q_{ii} + n \cdot q_{jj} \\
&= t + n \cdot q_{ii} \qquad (EM.2)
\end{aligned}$$

[22] Y de una columna, ya que la matriz $\mathbf{Q}$ es simétrica; entonces $\sum_{i=1}^{n} q_{ij} = 0$

donde $t = \sum_{i=1}^{n} q_{ii} = \text{tra}\,(\mathbf{Q})$. Y sumando por columnas:

$$\sum_{j=1}^{n} d_{ij}^2 = t + n \cdot q_{ii} \qquad (EM.3)$$

y sumando por filas $(EM.2)$ de nuevo:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 = 2\,n\,t \qquad (EM.4)$$

Sustituyendo en $(EM.1)$ la q_{jj} obtenida en $(EM.2)$ y q_{ii} en $(EM.3)$, se tiene:

$$d_{ij}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 - \frac{t}{n} + \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 - \frac{t}{n} - 2\,q_{ij}$$

y denotando $d_{i\cdot}^2 = \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2$ y $d_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2$ a las medias por filas y por columnas respectivamente, y utilizando $(EM.4)$, se obtiene:

$$d_{ij}^2 = d_{i\cdot}^2 + d_{\cdot j}^2 - d_{\cdot\cdot}^2 - 2\,q_{ij} \qquad (EM.5)$$

donde $d_{\cdot\cdot}^2$ es la media de todos los elementos de $\mathbf{D}$, dada por:

$$d_{\cdot\cdot}^2 = \frac{1}{n^2}\sum_{i}\sum_{j} d_{ij}^2$$

Finalmente, de $(EM.5)$ resulta que:

$$q_{ij} = -\frac{1}{2}\left(d_{ij}^2 - d_{i\cdot}^2 - d_{\cdot j}^2 + d_{\cdot\cdot}^2\right)$$

expresión que indica cómo construir la matriz de similaridad $\mathbf{Q}$, a partir de la matriz $\mathbf{D}$ de distancias.

Ahora, se considera el problema de obtener la matrix $\mathbf{X}$ dada la matriz $\mathbf{Q}$. Suponiendo que la matriz de similaridades es positiva definida de rango p, se obtiene su diagonalización y por ende puede representarse por:

$$\mathbf{Q} = \mathbf{V}\,\Lambda\,\mathbf{V}'$$

donde $\mathbf{V}$ es una matriz de $n \times p$ y Λ es diagonal[23], con elementos todos positivos[24]. En este sentido, se puede rescribir Λ como:

$$\Lambda = \Lambda^{1/2} \cdot \Lambda^{1/2}$$

Nótese entonces que:

$$\mathbf{Q} = \left(\mathbf{V}\,\Lambda^{1/2}\right)\left(\Lambda^{1/2}\,\mathbf{V}'\right) = \left(\mathbf{V}\,\Lambda^{1/2}\right)\left(\mathbf{V}\,\Lambda^{1/2}\right)' \qquad (EM.6)$$

y tomando:

$$\mathbf{Y} = \mathbf{V}\,\Lambda^{1/2}$$

se ha obtenido una matriz de $n \times p$ con p variables no correlacionadas que reproducen la métrica inicial. Notése que si se parte de ciertas variables, $\mathbf{X}_{n\times p}$; a partir de éstas se calculan las distancias con $(EM.0)$ y luego se aplica el método descrito a esta matriz de distancias, no se obtendrán las variables originales, $\mathbf{X}$, sino sus componentes principales. Esto es inevitable, ya que existe una indeterminación en el problema cuando la única información disponible son las distancias. En efecto, las distancias entre los elementos no varían si:

- Se modifican las medias de las variables[25].
- Se rotan los puntos, o equivalentemente, se multiplica por una matriz ortogonal.

Por $(EM.1)$, las distancias son función de los términos de la matriz de similaridad $\mathbf{Q}$, y esta matriz es invariante ante rotaciones de las variables. De hecho, si se define Θ=$\tilde{\mathbf{X}}\,\mathbf{A}$, donde $\mathbf{A}$ es ortogonal, entonces:

$$\Theta\,\Theta' = \left(\tilde{\mathbf{X}}\,\mathbf{A}\right)\left(\tilde{\mathbf{X}}\,\mathbf{A}\right)' = \tilde{\mathbf{X}}\,\left(\mathbf{A}\,\mathbf{A}'\right)\tilde{\mathbf{X}} = \tilde{\mathbf{X}}\,\tilde{\mathbf{X}}'$$

[23] De hecho, $\mathbf{V}$ contiene los vectores propios correspondientes a los valores propios no nulos de $\mathbf{Q}$; Λ es diagonal de $p \times p$ y contiene en su diagonal a dichos valores propios.

[24] Ésto debido a que $\mathbf{Q}$ es definida positiva.

[25] Se vió que $d_{ij}^2 = \sum_{s=1}^{p} (x_{is} - x_{js})^2 = \sum_{s=1}^{p} [(x_{is} - \bar{\mathbf{x}}_s) - (x_{js} - \bar{\mathbf{x}}_s)]^2$.

Ahora, la matriz $\mathbf{Q}$ sólo contiene "información" sobre el espacio generado por las variables $\mathbf{X}$. Cualquier rotación preserva las distancias y en consecuencia, cualquier rotación de las variables originales puede ser solución. Para verificar esto, sea $\mathbf{P}$ matriz de rotación (ortogonal) arbitraria, entonces de $(EM.6)$:

$$\begin{aligned}\mathbf{Q} &= \left(\mathbf{V}\Lambda^{1/2}\right)\left(\mathbf{P}\,\mathbf{P}'\right)\left(\mathbf{V}\Lambda^{1/2}\right)' \\ &= \left(\mathbf{V}\Lambda^{1/2}\mathbf{P}\right)\left(\mathbf{P}'\Lambda^{1/2}\mathbf{V}'\right) \\ &= \left(\mathbf{V}\Lambda^{1/2}\mathbf{P}\right)\left(\mathbf{V}\Lambda^{1/2}\mathbf{P}\right)'\end{aligned}$$

y en consecuencia $\mathbf{Y} = \left(\mathbf{V}\Lambda^{1/2}\,\mathbf{P}\right)$ es otra posible solución.

3.3.3.1 Obtención de las coordenadas principales.

Es frecuente que en la matriz de similitud obtenida a partir de la matriz de distancias, puedan identificarse los p valores propios más "importantes" en valor absoluto[26]. Si los restantes $n-p$ valores propios no nulos son mucho menores que los demás (y muy cercanos a cero en valor absoluto), se puede obtener una representación aproximada de los puntos utilizando los p vectores propios asociados a los primeros p valores propios positivos (ya ordenados) de la matriz de similaridad; en este caso las representaciones gráficas conservarán sólo aproximadamente la distancia entre los puntos.

Supóngase que se tiene una matriz de distancias, $\mathbf{D}$, al cuadrado. El procedimiento para obtener las *coordenadas principales* es:

1. Construir la matriz $\mathbf{Q} = -\frac{1}{2}\mathbf{P}\,\mathbf{D}\,\mathbf{P}$.

2. Obtener los valores propios de $\mathbf{Q}$ y tomar los r mayores valores propios,

[26] En el sentido de que son los mayores.

donde r se escoge de manera que los restantes $n-r$ valores propios sean próximos a cero. Dado que $\mathbf{P}\,\mathbf{1}=\mathbf{0}$, la matriz $\mathbf{Q}$ tiene rango máximo $n-1$, y siempre tendrá el vector propio $\mathbf{1}$ correspondiente al valor propio cero.

3. Obtener las coordenadas de los puntos en las variables mediante $\mathbf{v}_i\sqrt{\lambda_i}$, donde λ_i es un valor propio de $\mathbf{Q}$ y $\mathbf{v}_i$ su correspondiente vector propio. Esto implica aproximar $\mathbf{Q}$ por:

$$\mathbf{Q} \approx \left(\mathbf{V}_r\,\mathbf{\Lambda}_r^{1/2}\right)\left(\mathbf{\Lambda}_r^{1/2}\,\mathbf{V}_r'\right)$$

y tomar como coordenadas de los puntos las variables $\mathbf{Y}_r = \mathbf{V}_r \cdot \mathbf{\Lambda}_r^{1/2}$.

El método también aplicarse si la información de partida es directamente la matriz de similaridades, $\mathbf{Q}$, entre elementos. Como se ha visto, la *función de similaridad* es no negativa y simétrica. Si la matriz[27] de partida es precisamente $\mathbf{Q}$, entonces $q_{ii}=1$, $q_{ij}=q_{ji}$ y $0 \le q_{ij} \le 1$. Por lo tanto, la matriz de distancias asociadas será:

$$d_{ij}^2 = q_{ii}+q_{jj}-2q_{ij} = 2\,(1-q_{ij}) \qquad \Longleftrightarrow \qquad d_{ij} = \sqrt{2\,(1-q_{ij})}$$

y mediante un breve análisis se puede observar que $\sqrt{2\,(1-q_{ij})}$ es una distancia y que verifica la desigualdad del triángulo. Finalmente, pueden obtenerse medidas de la precisión conseguida mediante la aproximación, a partir de p valores propios positivos de la matriz de similaridad. Mardia ha propuesto el coeficiente:

$$m_{1,j} = 100 \times \frac{\sum_1^p \lambda_i}{\sum_1^n |\lambda_i|}$$

3.3.4. Relación entre coordenadas y componentes principales.

El escalamiento multidimensional representa un enfoque complementario a la técnica de las componentes principales en el sentido siguiente: componentes

[27] Matriz de similaridades.

principales considera la matriz de $p \times p$ de correlaciones (o covarianzas) entre variables, e investiga su estructura. El escalamiento multidimensional considera la matriz de $n \times n$ de distancias entre individuos e investiga su estructura. Ambos enfoques están claramente relacionados, y existen técnicas gráficas[28] que utlizan esta dualidad para representar conjuntamente las variables y los individuos en un mismo gráfico.

Cuando los datos originales están en la matriz $\tilde{\mathbf{X}}$ de individuos por variables y se construye la matriz $\mathbf{D}$ de distancias utilizando las distancias euclideanas entre los puntos de las variables originales, las coordenadas principales obtenidas de la matriz $\mathbf{D}$ son equivalentes a los componentes principales de las variables. En efecto, con variables de media cero, los componentes principales son los vectores propios de $\tilde{\mathbf{X}}' \tilde{\mathbf{X}}/n$, mientras que las coordenadas principales son los vectores propios estandarizados por $\sqrt{\lambda_i}$ de los valores propios de $\mathbf{Q} = \tilde{\mathbf{X}}' \tilde{\mathbf{X}}$. Además, por Álgebra Lineal básica, es claro que $\tilde{\mathbf{X}}' \tilde{\mathbf{X}}$ y $\tilde{\mathbf{X}} \tilde{\mathbf{X}}'$ tienen el mismo rango y los mismos valores propios no nulos.

Por otro lado, la matriz $n \times p$ que proporciona los valores de los p componentes principales en los n indivdiuos es $\mathbf{Z} = \tilde{\mathbf{X}} \mathbf{A}$, donde $\mathbf{Z}$ es $n \times p$ y tiene por columnas los componentes principales y $\mathbf{A}$ es $p \times p$ y contiene en columnas los vectores propios de $\tilde{\mathbf{X}}' \tilde{\mathbf{X}}$. La matriz $n \times p$ de coordenadas princiaples viene dada por:

$$\mathbf{Y} = [v_1, \ldots, v_p] \begin{bmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_p} \end{bmatrix} = \mathbf{V} \cdot \mathbf{L}$$

donde $\mathbf{v}_i$ es un vector propio de $\tilde{\mathbf{X}} \tilde{\mathbf{X}}'$, la matriz $\mathbf{V}$ es $n \times p$ y contiene los p vectores

[28] como la gráfica de biplot.

propios no nulos de $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$, y $\mathbf{L}$ es $p \times p$ es diagonal. Como $\mathbf{V}=\tilde{\mathbf{X}}$, es claro que, salvo por un factor de escala, ambos procedimientos conducen al mismo resultado.

El análisis en coordenadas principales o escalamiento multidimensional guarda una relación estrecha con componentes principales, sin embargo, el escalamiento multidimensional puede aplicarse a una gama más amplia de problemas, pues las coordenadas principales pueden obtenerse siempre, aunque las distancias de partida no hayan sido exactamente generadas a partir de variables, o bien, no se conozcan las variables originales y solo se cuente con una matriz de distancias.

3.3.5. Biplots.

Se conoce como *biplots* a las representaciones conjuntas en un plano de las filas y columnas de una matriz. En el caso de una matriz de datos, el biplot es una gráfico conjunto de las observaciones y las variables, la cual se obtiene a partir de la descomposición en valores singulares (SVD) de la matriz.

Una matriz $\mathbf{X}$ de dimensiones $n \times p$ puede siempre descomponerse como $\mathbf{X} = \mathbf{U}\mathbf{D}^{1/2}\mathbf{A}'$, donde $\mathbf{U}$ es $n \times p$ ortogonal y contiene en columnas los vectores propios asociados a valores propios no nulos de la matriz $\tilde{\mathbf{X}}\,\tilde{\mathbf{X}}'$, $\mathbf{D}$ es una matriz diagonal de orden p que contiene las raíces cuadradas de los valores propios no nulos de $\tilde{\mathbf{X}}'\,\tilde{\mathbf{X}}$ ó $\tilde{\mathbf{X}}\,\tilde{\mathbf{X}}'$ y $\mathbf{A}'$ es una matriz ortogonal de orden p y contiene, por filas, los vectores propios de $\tilde{\mathbf{X}}'\,\tilde{\mathbf{X}}$.

La descomposición en valores singulares tiene gran importancia práctica por que la mejor aproximación de rango $r < p$ a la matriz $\mathbf{X}$ se obtiene tomando los r valores propios mayores de $\tilde{\mathbf{X}}'\,\tilde{\mathbf{X}}$ y los correspondientes vectores propios de $\tilde{\mathbf{X}}'\,\tilde{\mathbf{X}}$

y construyendo:

$$\hat{\mathbf{X}} = \mathbf{U}_r \mathbf{D}_r^{1/2} \mathbf{A}_r'$$

donde $\mathbf{U}_r$ es $n \times r$ y contiene las primeras r columnas de $\mathbf{U}$ correspondientes a los r valores propios mayores de $\tilde{\mathbf{X}}\,\tilde{\mathbf{X}}'$; $\mathbf{D}_r^{1/2}$ es diagonal de $r \times r$ y contiene estos r valores propios; y $\mathbf{A}_r'$ de $r \times p$, contiene las r primeras filas de $\mathbf{A}'$ que corresponden a los r vectores propios de $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ ligados a los r valores propios mayores.

En términos prácticos, dicha representación consiste en aproximar $\mathbf{X}$ mediante la SVD de rango dos. Entonces, tomando $r = 2$, se tiene:

$$\mathbf{X} \approx \mathbf{U}_2\, \mathbf{D}_2^{1/2}\, \mathbf{A}_2' = \left(\mathbf{U}_2\, \mathbf{D}_2^{1/2-c/2}\right) \left(\mathbf{D}_2^{c/2}\, \mathbf{A}'\right) = \mathbf{F} \cdot \mathbf{C}$$

y escogiendo $0 \leq c \leq 1$ se obtienen distintas descomposiciones de $\mathbf{X}$ en dos matrices. La primera, $\mathbf{F}$, representa las n filas de la matriz $\mathbf{X}$ en un espacio de dos dimensiones y la segunda, $\mathbf{C}$, representa en el mismo espacio las columnas de la matriz. Según el valor de c se obtienen distintos biplots.

La precisión de la representación del biplot depende de la importancia de los dos primeros valores propios respecto al total. Si $(\lambda_1+\lambda_2)/\text{tra}(\mathbf{S})$, con $\mathbf{S} = \mathbf{ADA}'$, es próximo a uno, la representación será muy buena. Si este valor es pequeño (cercano a cero), el biplot no proporciona una representación fiable de los datos.

3.4. Análisis Discriminante.

El problema de *discriminación* o *clasificación* puede plantearse de varias formas y aparece en muchas áreas de la actividad humana. El planteamiento estadístico del problema es el siguiente: se dispone de un conjunto amplio de elementos que pueden venir de dos o más poblaciones distintas. Supongamos que en

cada elemento se ha observado una variable aleatoria p-dimensional $\mathbf{x}$, cuya distribución se conoce en las poblaciones consideradas. Se desea clasificar un nuevo elemento, con valores de las variables conocidas, en alguna de las poblaciones.

Existen varios enfoques posibles para este problema. Uno de éstos es el Análisis Discriminante clásico debido a Fisher, basado en el supuesto de que las matrices de varianzas y covarianzas son iguales, es decir, $\Sigma_i = \Sigma$; y que es óptimo bajo el supuesto de normalidad. Si todas las variables son continuas y ocurre que los datos no se siguen un distribución normal, es posible transformar las variables para obtener normalidad. Pero si se tienen variables continuas y discretas para clasificar, la hipótesis de normalidad multivariante no tendría sentido, sin embargo existen métodos que funcionan mejor en estos casos. Debido a la naturaleza de los datos a analizar en este proyecto, se expondrá el análisis discriminante robuzto clásico donde se supone una igualdad en las matrices de varianzas y covarianzas, pues de lo contrario, el número de parámetros a estimar sería elevado[29].

3.4.1. Planteamiento del problema.

Sean P_1 y P_2 dos poblaciones donde se ha definido una variable aleatoria p-dimensional, $\mathbf{x}$. Supóngase que $\mathbf{x}$ es absolutamente continua y que las funciones de densidad de ambas poblaciones, f_1 y f_2, son conocidas. Consideremos el problema de clasificar un nuevo elemento, $\mathbf{x}_0$, con valores conocidos de las p variables en una de estas poblaciones. Si se conocen las probabilidades *a priori*, digamos π_1, π_2, con $\pi_1 + \pi_2 = 1$, de que el elemento venga de cada una de las poblaciones, su distribución de probabilidad será:

$$f(\mathbf{x}) = \pi_1 f_1(\mathbf{x}) + \pi_2 f_2(\mathbf{2})$$

[29] Sin embargo, se incluirán algunas pruebas para la homoscedasticidad.

y una vez observado $\mathbf{x}_0$, se pueden calcular las probabilidades *a posteriori* de que el elemento haya sido generado por cada una de las dos poblaciones, $P(i|\mathbf{x}_0)$, con $i = 1, 2$. Estas probabilidades se calculan vía el teorema de Bayes:

$$P(i|\mathbf{x}_0) = \frac{P(\mathbf{x}_0|i)\pi_i}{\pi_1\, P(\mathbf{x}_0|1) + \pi_2\, P(\mathbf{x}_0|2)}$$

y como $P(\mathbf{x}_0|i) = f_i(\mathbf{x}_0)\mathbf{\Delta}\mathbf{x}_0$, se sigue que[30]:

$$P(1|\mathbf{x}_0) = \frac{f_1(\mathbf{x}_0)\,\pi_1}{f_1(\mathbf{x}_0)\,\pi_1 + f_2(\mathbf{x}_0)\,\pi_2}$$

y para la segunda población,

$$P(2|\mathbf{x}_0) = \frac{f_2(\mathbf{x}_0)\,\pi_2}{f_1(\mathbf{x}_0)\,\pi_1 + f_2(\mathbf{x}_0)\,\pi_2}$$

3.4.2. Opciones para decidir.

Respecto al caso previo, un criterio de clasificación es colocar a $\mathbf{x}_0$ en la población más probable *a posteriori*. Los denominadores son iguales, entonces se clasificará $\mathbf{x}_0$ en P_2 si:

$$\pi_2\, f_2(\mathbf{x}_0) > \pi_1\, f_1(\mathbf{x}_0)$$

Si las probabilidades *a priori* son iguales, la condición de clasificar en P_2 es:

$$f_2(\mathbf{x}_0) > f_1(\mathbf{x}_0)$$

es decir, se clasifica a $\mathbf{x}_0$ en la población más verosímil (o bien, más probable). Ahora bien, en un gran número de problemas de clasificación, los errores que se comenten tienen distintas consecuencias las cuales pueden cuantificarse. Por ejemplo, si una máquina automática clasifica equivocadamente un billete de \$20 como de \$50, al devolver un cambio equivocado el *coste* de clasificación será de

[30] Recuérdese que la densidad $f_i(\mathbf{x}_0)$ es proporcional a la probabilidad de que la observación $\mathbf{x}_0$ sea generada por la i-ésima población.

$30. En otros casos, estimar el coste puede ser más complejo, por ejemplo, si se clasifica un proceso productivo como en estado de control, la consecuencia (o *coste*) será una producción defectuosa, y si por error se detiene un proceso que funciona adecuadamente, el *coste* posiblemente será un *cuello de botella*.

En general, supóngase que las posibles decisiones en el problema de clasificación son únicamente dos: asignar en P_1 o P_2. Una regla de decisión es una partición del espacio muestral E_x en dos regiones A_1 y $A_2 = E_x - A_1$, tales que:

- Si $\mathbf{x}_0 \in A_1$, entonces clasificar en P_1.
- Si $\mathbf{x}_0 \in A_2$, entonces clasificar en P_2.

Si las consecuencias de un error de clasificación pueden cuantificarse, se pueden incluir en la solución del problema formulándolo como un problema bayesiano de decisión. Supóngase que:

1. Las consecuencias asociadas a los errores de clasificación son $c(2|1)$ y $c(1|2)$ donde $c(i|j)$ es el *coste* de clasificación en P_i de una unidad que pertenece a P_j. Estos costes se suponen conocidos.
2. El agente que decide, quiere maximizar su función de utilidad, lo cual equivale a minimizar el coste esperado.

Con estas dos hipótesis la mejor decisión es aquella que minimiza los costes esperados o funciones de pérdida de utilidad. Entonces, si se clasifica a un elemento en P_2, las posibles consecuencias son:

(a) Acertar, con probabilidad $P(2|\mathbf{x}_0)$, en cuyo caso no hay coste alguno de penalización;

(b) Equivocarse, con probabilidad $P(1|\mathbf{x}_0)$, en cuyo caso se incurrirá en el coste asociado $c(2|1)$.

El coste promedio[31], de la decisión "d_2: clasificar $\mathbf{x}_0$ en el P_2" será:

$$E(d_2) = c(2|1) \cdot P(1|\mathbf{x}_0) + 0 \cdot P(2|\mathbf{x}_0) = c(2|1) \cdot P(1|\mathbf{x}_0) \qquad (AD.1)$$

Análogamente, el coste esperado de la decisión "d_1: clasificar $\mathbf{x}_0$ en P_1" será:

$$E(d_1) = 0 \cdot P(1|\mathbf{x}_0) + c(1|2) \cdot P(2|\mathbf{x}_0) = c(1|2) \cdot P(2|\mathbf{x}_0) \qquad (AD.2)$$

El criterio de clasificación es asignar al elemento al grupo 2 si su coste esperado es menor, es decir, utilizando $(AD.1)$ y $(AD.2)$, si:

$$\frac{f_2(\mathbf{x}_0)\,\pi_2}{c(2|1)} > \frac{f_1(\mathbf{x}_0)\,\pi_1}{c(1|2)}$$

Esta condición indica que el elemento se clasificará en la población P_2 si:

(a) Su probabilidad *a priori* es más alta,

(b) La verosimilitud de que $\mathbf{x}_0$ provenga de P_2 es más alta,

(c) El coste de equivocarse al clasificarlo en P_2 es más bajo.

3.4.3. Poblaciones normales: función lineal discriminante.

Ahora, se aplicará el análisis anterior al caso en que f_1 y f_2 son disribuciones normales con distintos vectores de medias pero idéntica matriz de covarianzas. De manera general, supóngase que se desea clasificar un elemento genérico $\mathbf{x}$, que si pertenece a la población $i = 1, 2$ tiene la función de densidad:

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\,|V|^{1/2}} \exp\left\{-\frac{1}{2}\,(\mathbf{x} - \mu_i)'\,V^{-1}\,(\mathbf{x} - \mu_i)\right\}$$

La partición óptima es, de acuerdo con la sección anterior, *clasificar* en la población P_2 si:

$$\frac{f_2(\mathbf{x})\,\pi_2}{c(2|1)} > \frac{f_1(\mathbf{x})\,\pi_1}{c(1|2)}$$

[31] Valor esperado o coste esperado.

Como ambos términos son siempre positivos, tómese logaritmos y sustituyendo $f_i(\mathbf{x})$ por su expresión, la ecuación anterior se convierte en:

$$-\frac{1}{2}(\mathbf{x}-\mu_2)' V^{-1}(\mathbf{x}-\mu_2) + \log\frac{\pi_2}{c(2|1)} > -\frac{1}{2}(\mathbf{x}-\mu_1)' V^{-1}(\mathbf{x}-\mu_1) + \log\frac{\pi_1}{c(1|2)}$$

Si se denota D_i^2 a la distancia de Mahalanobis entre el punto observado, $\mathbf{x}$, y la media de la i-ésima población:

$$D_i^2 = (\mathbf{x}-\mu_i)' V^{-1}(\mathbf{x}-\mu_i)$$

se puede escribir:

$$D_1^2 - \log\frac{\pi_1}{c(1|2)} > D_2^2 - \log\frac{\pi_2}{c(2|1)}$$

y suponiendo igualdad de costes y de las probabilidades *a priori*, c(1|2)=c(2|1); $\pi_1 = \pi_2$, la regla anterior se reduce a:

Clasificar en P_2 si $D_1^2 > D_2^2$ ... (AD.3)

es decir, clasificar la observación en la población de cuya media esté más próxima, midiendo la distancia con la medida de Mahalanobis. Obsérvese que si las variables $\mathbf{x}$ tuvieran como matriz de varianzas y covarianzas $V = \mathbf{I}\sigma^2$, la regla (AD.3) equivaldría a utilizar la distancia euclideana.

3.4.4. Generalización para varias poblaciones normales.

La generalización para G poblaciones es de la siguiente manera: el objetivo es ahora dividir el espacio E_x en G regiones $A_1, \ldots, A_g, \ldots, A_G$ tales que si $\mathbf{x}$ pertenece a A_g, el punto se clasifica en la población P_g. Suponiendo que los costes de clasificación son constantes y no dependen de la población en que se haya clasificado, entonces la región A_g quedará definida por aquellos puntos con máxima probabilidad de ser generados por P_g, es decir, donde el producto de la

probabilidad *a priori* y la verosimilitud sean máximas:

$$A_g = \{\mathbf{x} \in E_x | \pi_g \cdot f_g(\mathbf{x}) > \pi_i \cdot f_i(\mathbf{x}); \forall i \neq g\} \qquad (AD.4)$$

Si las probabilidades a priori son iguales, entonces $\pi_i = G^{-1}\ \forall i$, y las distribuciones $f_i(\mathbf{x})$ son normales con la misma matriz de varianzas, la condición $(AD.4)$ equivale a calcular la distancia de Mahalanobis del punto observado al centro de cada población y clasificarle en la población que minimice esta distancia. Ahora, minimizar las distancias de Mahalanobis $(\mathbf{x} - \mu_g)' V^{-1} (\mathbf{x} - \mu_g)$ equivale, eliminando el término $\mathbf{x}' V^{-1} \mathbf{x}$ que aparece en todas las ecuaciones, a minimizar el operador lineal:

$$L_g(\mathbf{x}) = -2\mu_g' V^{-1} \mathbf{x} + \mu_g' V^{-1} \mu_g \qquad (AD.5)$$

y llamando $\mathbf{w}_g = V^{-1}\mu_g$ la regla es:

$$\min_g \left(\mathbf{w}_g' \mu_g - 2\mathbf{w}_g' \mathbf{x} \right)$$

Para interpretar esta regla, nótese que la *frontera* de separación entre dos poblaciones, (i, j), vendrá definida por:

$$A_{ij}(\mathbf{x}) = L_i(\mathbf{x}) - L_j(\mathbf{x}) = 0 \qquad (AD.6)$$

Sustituyendo con $(AD,5)$ y reordenando los términos se obtiene:

$$A_{ij}(\mathbf{x}) = 2 (\mu_i - \mu_j)' V^{-1} \mathbf{x} + (\mu_i - \mu_j)' V^{-1} (\mu_i + \mu_j) = 0$$

y denotando $\mathbf{w}_{ij} = V^{-1} (\mu_i - \mu_j) = \mathbf{w}_i - \mathbf{w}_j$, la *frontera* puede escribirse como:

$$\mathbf{w}_{ij}' \cdot \mathbf{x} = \mathbf{w}_{ij}' \cdot \left[\frac{1}{2} (\mu_i + \mu_j) \right]$$

Observación.

Es importante resaltar que la regla de decisión obtenida cumple la propiedad transitiva. Por ejemplo, si $G = 3$, y obtenemos que para un punto $\mathbf{x}$:

$$D_1^2(\mathbf{x}) > D_2^2(\mathbf{x}) \qquad y \qquad D_2^2(\mathbf{x}) > D_3^2(\mathbf{x})$$

entonces, se concluye que $D_1^2(\mathbf{x}) > D_3^2(\mathbf{x})$ el cual será el resultado que se obtiene si se calculan estas distancias. Además, si $p = 2$, cada una de las tres ecuaciones $A_{ij}(\mathbf{x}) = 0$, será una recta y las tres se cortarán en el mismo punto. En efecto, cualquier otra recta que pase por el punto de corte de las rectas $A_{12}(\mathbf{x}) = 0$ y $A_{23}(\mathbf{x}) = 0$ tiene la expresión:

$$a_1 A_{12}(\mathbf{x}) + a_2 A_{23}(\mathbf{x}) = 0$$

ya que si $\mathbf{x_0^*}$ es el punto de corte, se sigue que $A_{12}(\mathbf{x}^*) = 0$, por pertenecer a la primera recta y que $A_{23}(\mathbf{x}^*) = 0$ por pertenecer a la segunda, y por tanto:

$$a_1 A_{12}(\mathbf{x}^*) + a_2 A_{23}(\mathbf{x}^*) = a_1 \cdot 0 + a_2 \cdot 0 = 0$$

Ahora, por $(AD.6)$, $A_{13}(\mathbf{x}) = L_1(\mathbf{x}) - L_3(\mathbf{x}) = L_1(\mathbf{x}) - L_2(\mathbf{x}) + L_2(\mathbf{x}) - L_3(\mathbf{x})$, y se tiene que:

$$A_{13}(\mathbf{x}) = A_{12}(\mathbf{x}) + A_{23}(\mathbf{x})$$

así, la recta $A_{13}(\mathbf{x})$ debe siempre pasar por el punto de corte de las otras dos.

3.4.5. Poblaciones desconocidas: caso general.

En esta sección se exhibirá la teoría cuando en lugar de trabajar con poblaciones, se dispone de muestras. Se abordará el caso de G poblaciones posibles. La matriz general de datos $\mathbf{X}$, de dimensiones $(n \times p)$, puede considerarse ahora particionada en G matrices, correspondientes a las subpoblaciones. Llámese x_{ijg} a los elementos de estas submatrices, donde i representa el individuo, j la variable y g el grupo o submatriz. Denótese n_g al número de elementos en el grupo g, entonces el número total de observaciones es:

$$n = \sum_{g=1}^{G} n_g$$

Llámese $\mathbf{x}'_{ig}$ al vector fila $(1 \times p)$ que contiene los p valores de las variables para el i-ésimo individuo en el grupo g, es decir, $\mathbf{x}'_{ig} = (x_{i1g}, x_{i2g}, \ldots, x_{ipg})$. El vector de medias dentro de cada clase o subpoblación será:

$$\bar{\mathbf{x}}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} \mathbf{x}_{ig}$$

que es un vector columna de dimensión p que contiene las p medias para las observaciones de la clase g. La matriz de varianzas y covarianzas estimada para los elementos de la clase g será:

$$\hat{\mathbf{S}}_g = \frac{1}{n_g - 1} \sum_{i=1}^{n_g} (\mathbf{x}_{ig} - \bar{\mathbf{x}}_g) (\mathbf{x}_{ig} - \bar{x}_{ig})'$$

Si se supone que las G poblaciones tienen la misma matriz de varianzas y covarianzas, su mejor estimación centrada con todos los datos será una combinación lineal de las estimaciones centradas de las matrices de covarianza en cada población, con peso proporcional a su precisión. Por tanto:

$$\hat{\mathbf{S}}_w = \sum_{g=1}^{G} \frac{n_g - 1}{n - G} \hat{\mathbf{S}}_g$$

Para obtener las funciones discriminantes, se utilizará $\bar{\mathbf{x}}_g$ como estimación de μ_g y $\hat{\mathbf{S}}_w$ como estimación de $\mathbf{V}$. Entonces, suponiendo probabilidades *a priori* y los costes de clasificación iguales, se clasificará al elemento en el grupo que conduzca a un valor mínimo de la distancia de Mahalanobis entre el punto $\mathbf{x}$ y la media del grupo. Es decir, denotando $\hat{\mathbf{w}}_g = \hat{\mathbf{S}}^{-1} \bar{\mathbf{x}}_g$ se clasificará un nuevo elemento $\mathbf{x}_0$ en aquella población g donde:

$$\min_g (\mathbf{x}_0 - \bar{\mathbf{x}}_g)' \hat{\mathbf{S}}_w^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_g) = \min_g \left[\hat{\mathbf{w}}'_g (\bar{\mathbf{x}}_g - \mathbf{x}_0) \right]$$

que equivale a construir las variables indicadoras escalares $z_{g,g+1} = \hat{\mathbf{w}}'_g \mathbf{x}_0$, para $g = 1, 2, \ldots, G$, donde:

$$\hat{\mathbf{w}}_{g,g+1} = \hat{\mathbf{S}}_w^{-1} (\bar{\mathbf{x}}_g - \bar{\mathbf{x}}_{g+1}) = \hat{\mathbf{w}}_g - \hat{\mathbf{w}}_{g+1}$$

y clasificar en la población g o en la población $g+1$ si:

$$|z_{g,g+1} - \hat{m}_g| < |z_{g,g+1} - \hat{m}_{g+1}|$$

donde $\hat{m}_g = \hat{\mathbf{w}}'_{g,g+1} \cdot \bar{\mathbf{x}}_g$. En el caso particular de dos grupos, la fución discriminante lineal $\hat{w} = \hat{\mathbf{S}}_w^{-1} (\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1)$ puede obtenerse por regresión, definiendo una variable ficticia que tome los valores cero o uno según el dato pertenezca a una u otra población.

3.4.6. Discriminación cuadrática: discriminación en poblaciones no normales.

Si admitiendo la normalidad de las observaciones, la hipótesis de igualdad de varianzas no fuese admisible, el procedimiento para resolver el problema de discriminación es clasificar la observación en el grupo con máxima probabilidad *a posteriori*. Esto equivale a clasificar la observación $\mathbf{x}_0$ en el grupo donde se minimice la función:

$$\min_{j \in (1,\ldots,G)} \left[\frac{1}{2} \log |\mathbf{V}_j| + \frac{1}{2} (\mathbf{x}_0 - \mu_j)' \mathbf{V}'_j (\mathbf{x}_0 - \mu_j) - \log (C_j \cdot \pi_j) \right]$$

Cuando $\mathbf{V}_j$ y μ_j son desconocidos, se estiman por $\mathbf{S}_j$ y $\bar{\mathbf{x}}_j$ de la forma habitual. Ahora, el término $\mathbf{x}'_0 \mathbf{V}'_j \mathbf{x}_0$ no puede anularse, pues depende del grupo, y las funciones discriminantes no son lineales y tendrán un término de segundo grado. Suponiendo que los costes de clasificación son iguales en todos los grupos, se clasificarán nuevas observaciones con la regla:

$$\min_{j \in (1,\ldots,G)} \left[\frac{1}{2} \log |\hat{\mathbf{V}}_j| + \frac{1}{2} (\mathbf{x}_0 - \hat{\mu}_j)' \hat{\mathbf{V}}'_j (\mathbf{x}_0 - \hat{\mu}_j) - \log (\pi_j) \right]$$

en el caso particular de dos poblaciones y suponiendo las mismas probabilidades *a priori*, se clasificará una nueva observación en la población dos, si:

$$\log |\hat{\mathbf{V}}_1| + (\mathbf{x}_0 - \hat{\mu}_1)' \hat{\mathbf{V}}'_1 (\mathbf{x}_0 - \hat{\mu}_1) > \log |\hat{\mathbf{V}}_2| + (\mathbf{x}_0 - \hat{\mu}_2)' \hat{\mathbf{V}}'_2 (\mathbf{x}_0 - \hat{\mu}_2)$$

que equivale a:

$$\mathbf{x}_0' \left(\hat{\mathbf{V}}_1^{-1} - \hat{\mathbf{V}}_2^{-1} \right) \mathbf{x}_0 - 2\,\mathbf{x_0}' \left(\hat{\mathbf{V}}_1^{-1}\,\hat{\mu}_1 - \hat{\mathbf{V}}_2^{-1}\,\hat{\mu}_2 \right) > c \qquad (AD.7)$$

donde:

$$c = \log \left(|\hat{\mathbf{V}}_2| / |\hat{\mathbf{V}}_1| \right) + \hat{\mu}_2'\,\hat{\mathbf{V}}_2^{-1}\,\hat{\mu}_2 - \hat{\mu_1}'\,\hat{\mathbf{V}}_1^{-1}\,\hat{\mu}_1$$

Ahora, denotando:

$$\hat{\mathbf{V}}_d^{-1} = \left(\hat{\mathbf{V}}_1^{-1} - \hat{\mathbf{V}}_2^{-1} \right) \qquad y \qquad \hat{\mu}_d = \hat{\mathbf{V}}_d \left(\hat{\mathbf{V}}_1^{-1}\,\mu_1 - \hat{\mathbf{V}}_2^{-1}\,\mu_2 \right)$$

definiendo las nuevas variables:

$$\mathbf{z}_0 = \hat{\mathbf{V}}_d^{-1/2} \cdot \mathbf{x}_0$$

y llamando $\mathbf{z}_0 = (z_{01}, \ldots, z_{0p})'$ y $\mathbf{m} = (m_1, \ldots, m_p)' = \hat{\mathbf{V}}_d^{(1/2)} \left(\hat{\mathbf{V}}_1^{-1}\,\mu_1 - \hat{\mathbf{V}}_2^{-1}\,\mu_2 \right)$, la ecuación $(AD.7)$ puede escribirse como:

$$\sum_{i=1}^{p} z_{0i}^2 - 2 \sum_{i=1}^{p} z_{0i}\,m_i > c$$

Ésta es una ecuación de segundo grado en las nuevas variables z_{0i}. Las regiones resultantes con estas funciones de segundo grado son típicamente disjuntas y difíciles de interpretar en varias dimensiones. El número de parámetros a estimar en el caso cuadrático es mucho mayor que en el caso lineal. En el caso lineal hay que estimar $Gp + p(p+1)2$ y en el cuadrático $G(p + p(p+1)2)$. Por ejemplo, con 10 variables y 4 grupos se pasa de estimar 95 parámetros en el caso lineal a 260 en el caso cuadrático. Este gran número de parámetros hace que, excepto para muestras muy grandes, la discriminación cuadrática sea bastante inestable y, aunque las matrices de covarianzas sean muy diferentes, se obtengan con frecuencia mejores resultados con la función lineal que con la cuadrática.

Un problema adicional con la función discriminante cuadrática es que es muy sensible a desviaciones de la normalidad de los datos. La evidencia disponible indica que la clasificación lineal es más robusta.

La discriminación cuadrática aparece también en el análisis de determinadas poblaciones no normales, como por ejemplo Lachenbruch(1975). Para poblaciones arbitrarias se tienen dos alternativas:

(a) Aplicar la teoría general expuesta en **3.4.2** y obtener una función discriminante complicada,

(b) Aplicar la teoría de poblaciones normales, tomar como medida de distancia la de Mahalanobis y clasificar $\mathbf{x}$ en la población P_j para la cual $\mathbf{D}^2$:

$$\mathbf{D}^2 = (\mathbf{x} - \bar{\mathbf{x}}_j)' \, \hat{\mathbf{V}}_j^{-1} \, (\mathbf{x} - \bar{\mathbf{x}}_j)$$

sea mínima.

En el caso particular de querer discriminar *mezclas* de variables (continuas y discretas) o bien, cuando no se tiene la igualdad estadística en las matrices de varianzas y covarianzas, $\Sigma_i \neq \Sigma_j$ para algún par (i, j), existen métodos alternativos tales como la Regresión Logística y/o la Regresión Logística Multinomial.

3.5. Comentarios sobre el proyecto.

Las técnicas multivariadas previamente descritas, entre otras, serán utilizadas en el desarrollo de este proyecto. No se debe perder de vista que el objetivo es proponer una regionalización de los Estados Unidos Mexicanos con base a indicadores electorales, como alternativa a la regionalización propuesta por CONAPO con base en el Índice de Marginación.

Se considerarán los estados como los "individuos" y los indicadores como variables. Mediante la técnica de componentes principales, así como la gráfica de los dos primeros componentes, se espera obtener grupos de estados que sean homogéneos y representados por variables no correlacionadas. De manera análoga se procederá con el análisis de conglomerados mediante los distintos algoritmos jerárquicos y los dendrogramas. Con las gráficas Biplot, se pretende verificar el comportamiento de los estados e indicadores en conjunto. El análisis discriminante se utilizará para validar las regiones propuestas, y para presentar pruebas estadísticas multivariadas. Éste último se realizará con el software S-PLUS V6.1.

Finalmente, para considerar la matriz de distancias en la etapa de aplicación de estas técnicas, se utilizarán las distintas opciones que proporciona el software STATA 9.2, con la intención de corroborar y principalmente, comparar los resultados obtenidos. Las técnicas adicionales que se utilizarán serán descritas con detalle en los Apéndices.

Capítulo 4

Resultados de la Aplicación de las Técnicas.

En este capítulo se presentan los resultados así como las regiones obtenidas con las distintas técnicas aplicadas. Se utilizó el software STATA versión 9.2, y los análisis se llevaron a cabo con base en las especificaciones del capítulo anterior. Como ya se mencionó, se consideraron indicadores de las Verificaciones Nacionales Muestrales de los años 2005 y 2006, y fueron los siguentes [1]:

- **Encuesta de Cobertura.**

  1. Empadronados.
  2. Empadronados en el estado.
  3. Empadronados en la sección.
  4. Credencializados.
  5. Credencializados en el estado.
  6. Credencializados en la sección.

[1] Se trabajó con los porcentajes de cada indicador por entidad federativa, pues son los datos que proporciona el IFE.

- **Encuesta de Actualización.**

  1. Fallecidos en el padrón.
  2. Cambios de domicilio no reportados en el padrón.
  3. Cambios de domicilio no reportados al mismo municipio.
  4. Cambios de domicilio no reportados a otro municipio dentro del mismo estado.
  5. Cambios de domicilio no reportados a otro estado.
  6. Cambios de domicilio no reportados a otro país.

Es importante mencionar que dichos indicadores no representan al total de éstos en las encuestas; en efecto, se optó por considerar a los ya mencionados, debido a que históricamente han sido los más utilizados y por ende, mantienen una presencia dentro de ambas encuestas. Se debe añadir el hecho de que los indicadores excluidos no siempre se han medido, o bien, su inclusión en las encuestas no ha sido constante, y en ese sentido, su influencia sería mínima o dificil de determinar en lo que respecta a representar la calidad del padrón. En el Apéndice C se encuentran tanto las definiciones de los indicadores utilizados en estos análisis, así como las fórmulas aplicadas para estimar su valor numérico.

Con la intención de facilitar la notación, se utilizará la siguiente simbología:

Aguascalientes ... 1		Guerrero ... 12		Quintana Roo ... 23
Baja California ... 2		Hidalgo ... 13		San Luis Potosí... 24
Baja California Sur ... 3		Jalisco ... 14		Sinaloa ... 25
Campeche ... 4		México ... 15		Sonora ... 26
Coahuila ... 5		Michoacán ... 16		Tabasco ... 27
Colima ... 6		Morelos ... 17		Tamaulipas ... 28

Chiapas ... 7		Nayarit ... 18		Tlaxcala ... 29
Chihuahua ... 8		Nuevo León ... 19		Veracruz ... 30
Distrito Federal ... 9		Oaxaca ... 20		Yucatán ... 31
Durango ... 10		Puebla ... 21		Zacatecas ... 32
Guanajuato ... 11		Querétaro ... 22		

4.1. Resultados obtenidos con la VNM2005.

En primera instancia se presentarán los resultados obtenidos con la VNM2005; por ello, en las subsecciones siguentes hasta antes de la sección 4.2, se exhibirán:

1. Los dendrogramas resultantes del Análisis de Conglomerados y las regiones formadas al proceder con los cortes correpondientes. Es importante mencionar que es posible construir un número mayor o menor de clusters dependiendo de la altura del corte; sin embargo, se optó por mantener la misma estructura en las regiones que por realizar los cortes a una misma altura.

2. Las regiones obtenidas con el Algoritmo de Partición k-medias para cuatro y cinco clusters. Cabe mencionar que los resultados del Algoritmo de Partición de las k-medias se presentan para cuatro y cinco clusters pues se optó por mantener la estructura obtenida en los dendrogramas. En éstos, como se verá a continuación, el número de regiones obtenidas varió, precisamente, de cuatro a cinco.

3. Gráfica de los dos primeros componentes principales obtenidos con la matriz de varianzas y covarianzas.

4. Gráfica de los dos primeros componentes principales obtenidos con la matriz de correlaciones.

5. Biplots suponiendo homoscedasticidad.

Respecto al Algortimo de Partición de las K - medias, cabe mencionar que el software STATA V9.2 presenta distintas opciones para los *centros iniciales de grupo*. Éstas son:

1. **Tomar k observaciones aleatorias.** Esta es la modalidad que el software guarda por *default* y especifica que k observaciones únicas son elegidas de manera aleatoria de entre aquellas que serán agrupadas.

2. **Tomar las primeras k observaciones.** Las primeras k observaciones de entre todas las que serán agrupadas, se utilizan como centros iniciales de los k grupos.

3. **Tomar las últimas k observaciones.** Las últimas k observaciones de entre todas las que serán agrupadas, se utilizan como centros iniciales de los k grupos.

4. **Tomar k centros aleatorios de grupos dentro del rango de datos.** En esta modalidad se generan de manera aleatoria los k centros iniciales de los grupos. Estos valores son seleccionados de manera aleatoria, asumiendo que los datos siguen una distribución uniforme sobre su rango.

5. **Grupos de medias de k particiones aleatorias de los datos.** Se forman k particiones aleatorias entre las observaciones que serán agrupadas. El grupo de medias que provienen de los k grupos definidos en dicha partición son utilizados como centros iniciales de grupos.

6. **Grupos de medias de k particiones formados agrupando cada k-ésima observación.** En esta modalidad, se forman k particiones asignando las observaciones 1, $1 + k$, $1 + 2k$, ... al primer grupo; las observaciones 2,$2 + k$, $2 + 2k$, ... al segundo grupo; y así sucesivamente hasta formar los k grupos. Las medias de éstos se utlizan como centros inciales.

7. **Grupos de medias de k particiones contiguas (igualmente espaciadas) de los datos.** Se forman k particiones igualmente espaciadas de los datos, donde aproximadamente las primeras N/k observaciones se asignan a la primera partición, las siguientes N/k observaciones se asignan a la segunda, y así sucesivamente. Las medias generadas se utilizan como centros iniciales de los grupos.

8. **Grupos de medias formadas de particiones definidas por una variable inicial que agrupe.** Se debe proporcionar una variable *agrupadora* que defina los k grupos entre las observaciones que serán agrupadas. Las medias de estos k grupos se utilizan como centros iniciales de los grupos.

Asimismo, el software presenta distintas opciones como medidas de similaridad (y disimilaridad) como son la medida euclideana, la medida L1,la matriz de correlaciones, etc. Para todos los análisis [2] se exhibirán dos casos: los resultados con la norma euclideana y los resultados tomando la matriz de correlaciones.

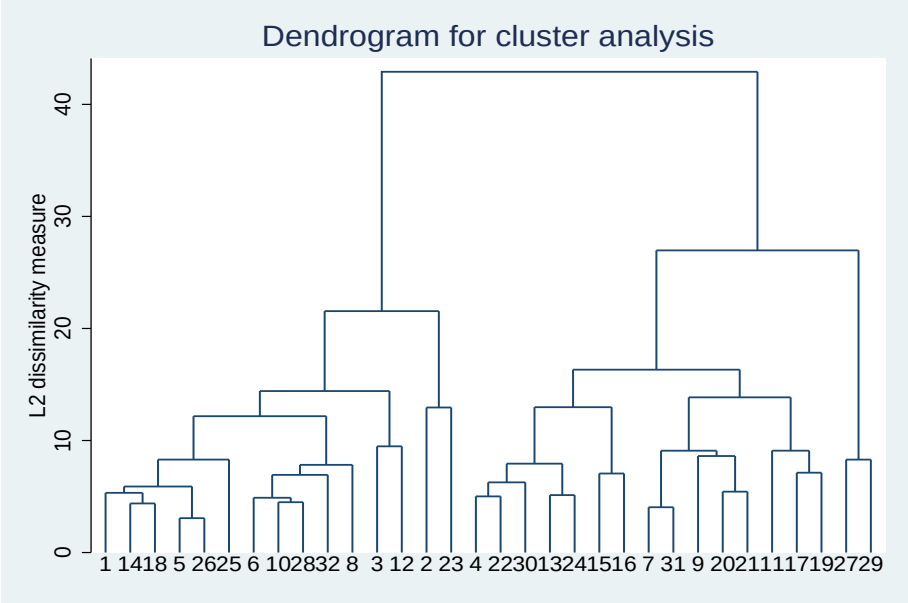
Cabe mencionar que los resultados que se presentarán en todos los análisis con el Algoritmo de Partición de las k - medias, se obtuvieron bajo la primera modalidad, es decir, tomando k observaciones aleatorias de los datos como centros iniciales de grupos. Esto fue debido principalmente a que las observaciones no mantienen propiedades distribucionales específicas. Adicionalmente, se realizaron corridas del Algortimo de Partición K-medias con las siguientes tres modalidades descritas que proporciona el software [3], y esencialmente se obtuvieron las mismas regiones que se presentan como propuesta final de estos análisis.

[2] Estos análisis incluyen los resultados con ambas Verificaciones y algunas combinaciones lineales de indicadores consideradas, como el promedio de los indicadores, etc.

[3] Tomando como centros inciales de grupo las primeras k observaciones, las últimas k observaciones y tomando k centros aleatorios de grupos dentro del rango de datos.

4.1.1. Análisis de Conglomerados.

4.1.1.1. Análisis de Conglomerados con *liga completa*, tomando como medida de disimilaridad la norma euclideana.

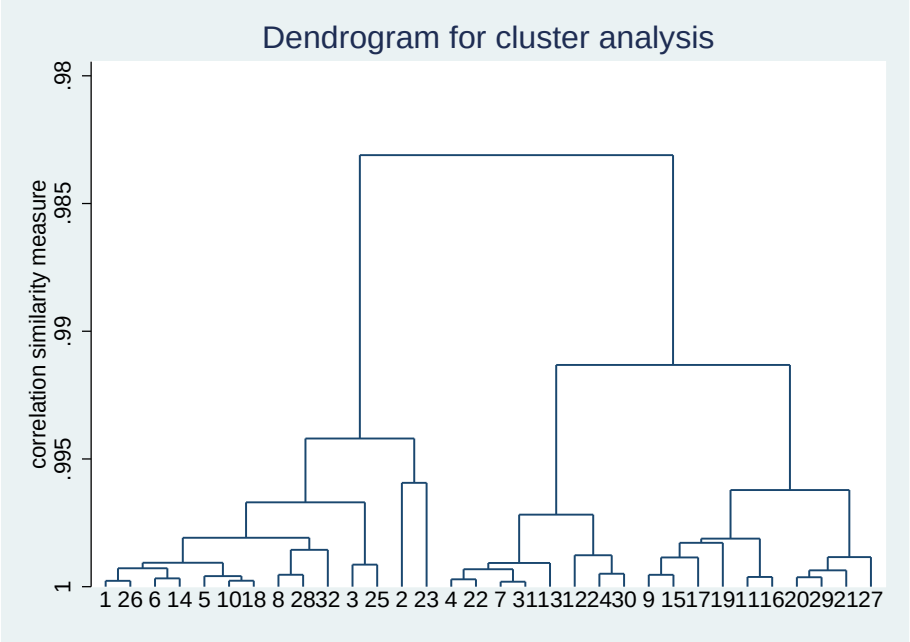


Realizando el corte correspondiente se obtienen las siguientes regiones:

Región 1	(Aguascalientes, Jalisco, Nayarit, Coahuila, Sonora, Sinaloa, Colima, Durango, Tamaulipas, Zacatecas, Chihuahua, Baja California Sur, Guerrero.
Región 2	Baja California, Quintana Roo).
Región 3	(Campeche, Querétaro, Veracruz, Hidalgo, San Luis Potosí, México, Michocán.
Región 4	Chiapas, Yucatán, Distrito Federal, Oaxaca, Puebla, Guanajuato, Morelos, Nuevo León).
Región 5	Tabasco, Tlaxcala.

Nótese que desde un principio, los estados de Baja California (2) y Quintana Roo (23) se "separan" del resto formando una sola región. La misma situación ocurre con los estados de Tabasco y Tlaxcala. Se presenta un efecto *encadenamiento* en el dendrograma, e incluso considerarse solamente tres regiones.

4.1.1.2. Análisis de Conglomerados con *liga completa*, tomando como medida de similaridad la matriz de correlaciones.

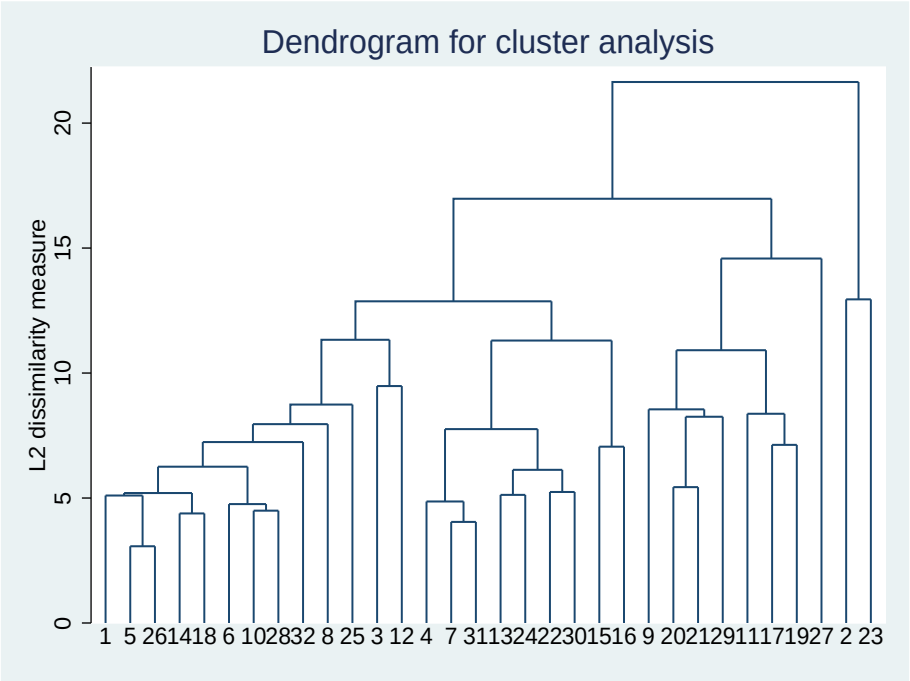


Al realizar el corte correspondiente respecto a la medida de similaridad, se obtienen las siguientes regiones:

Región 1	(Aguascalientes, Sonora, Colima, Jalisco, Coahuila, Durango, Nayarit, Chihuahua, Tamaulipas, Zacatecas, Baja California Sur, Sinaloa.
Región 2	Baja California, Quintana Roo).
Región 3	(Campeche, Querétaro, Chiapas, Yucatán, Hidalgo, Guerrero, San Luis Potosí, Veracruz.
Región 4	Distrito Federal, México, Morelos, Nuevo León, Guanajuato, Michoacán).
Región 5	Oaxaca, Tlaxcala, Puebla, Tabasco.

Nótese la similitud que guardan estas regiones respecto a las obtenidas en el caso anterior. Las regiones 1, 2 y 3 en esencia, se mantienen; mientras que en las regiones 4 y 5 los cambios son menores. Incluso, con base en el dendrograma, se podrían formar dos o tres regiones, como se muestra con los paréntesis.

4.1.1.3. Análisis de Conglomerados con *liga promedio*, tomando como medida de disimilaridad la norma euclideana.

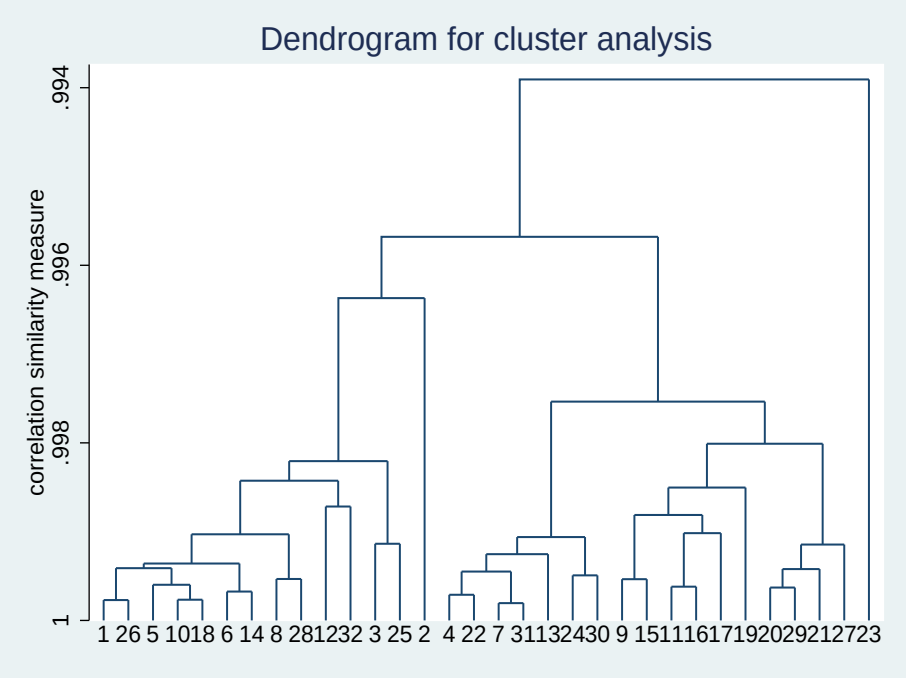


Al ejecutar el corte correspondiente respecto a la medida de disimilaridad, resultan las siguientes regiones:

Región 1	(Aguascalientes, Coahuila, Sonora, Jalisco, Nayarit, Colima, Durango, Tamaulipas, Zacatecas, Chihuahua, Sinaloa, Baja California Sur, Guerrero.
Región 2	Campeche, Chiapas, Yucatán, Hidalgo, San Luis Potosí, Querétaro, Veracruz, México, Michocán).
Región 3	Distrito Federal, Oaxaca, Puebla, Tlaxcala, Guanajuato, Morelos, Nuevo León, Tabasco.
Región 4	Baja California, Quintana Roo.

La región 1 se mantiene respecto a los casos anteriores. Nótese también que si a la **Región** 3 se le une Chiapas y Yucatán, resultan regiones análogas a las obtenidas. Se presenta el efecto de *encadenamiento* del dendrogama, lo cual puede generar solo tres regiones, como muestran los paréntesis.

4.1.1.4. Análisis de Conglomerados con *liga promedio*, tomando como medida de similaridad la matriz de correlaciones.

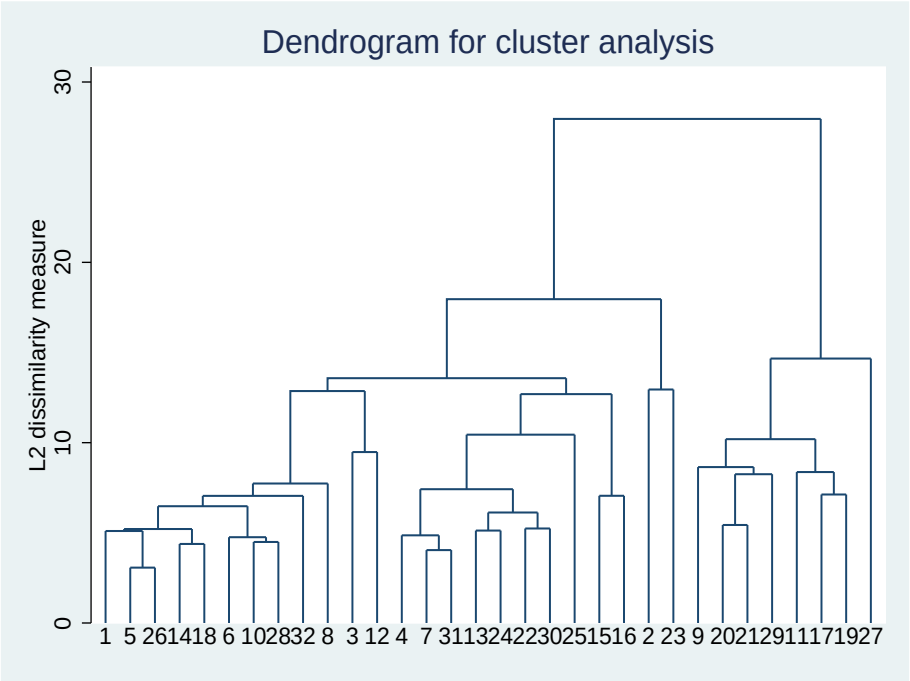


Con el corte correspondiente, se obtienen las siguientes regiones:

Región 1	(Aguascalientes, Sonora, Coahuila, Durango, Nayarit, Colima, Jalisco, Chihuahua, Tamaulipas, Guerrero, Zacatecas, Baja California Sur, Sinaloa).
Región 2	Baja California.
Región 3	(Campeche, Querétaro, Chiapas, Yucatán, Hidalgo, San Luis Potosí, Veracruz).
Región 4	(Distrito Federal, México, Guanajuato, Michoacán, Morelos, Nuevo León.
Región 5	Oaxaca, Tlaxcala, Puebla, Tabasco).
Región 6	Quintana Roo.

Nótese que las regiones 1, 3 y 4 prácticamente preservan su estructura. El "cambio" se presenta con los estados de Baja California y Quintana Roo al separarse. Pero con base al dendrograma, pueden formarse cinco regiones, donde las regiones 2 y 5 solo estarán formadas por un estado (Baja California y Quintana Roo respectivamente).

4.1.1.5. Análisis de Conglomerados con *liga peso - promedio ponderado* (Media de grupos), tomando como medida de disimilaridad la norma euclideana.

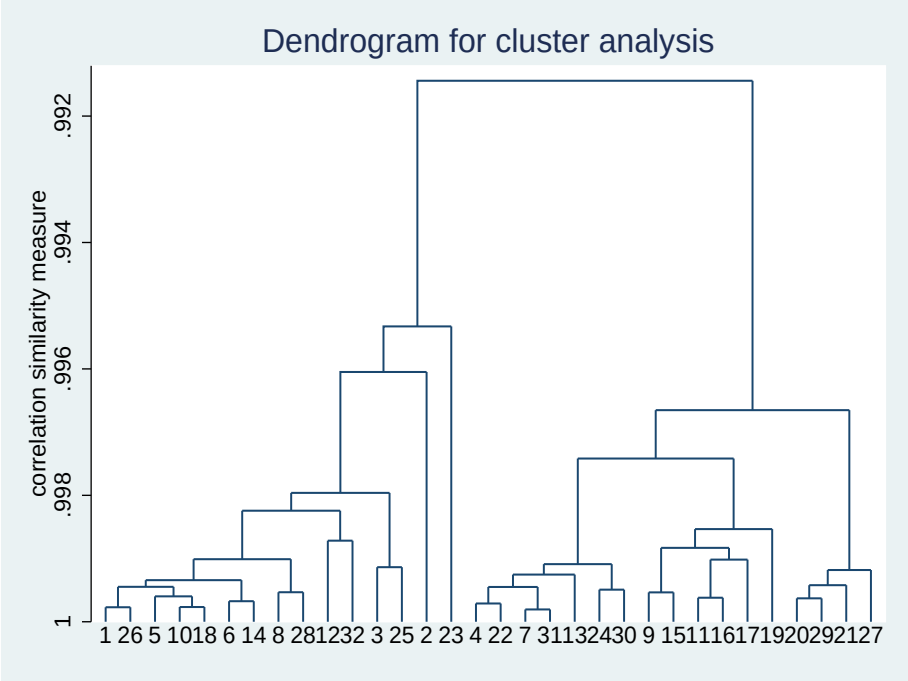


Al realizar el corte correspondiente respecto a la medida de disimilaridad, se obtienen las siguientes regiones:

Región 1	Aguascalientes, Coahuila, Sonora, Jalisco, Nayarit, Colima, Durango, Tamaulipas, Zacatecas, Chihuahua, Baja California Sur, Guerrero.
Región 2	Campeche, Chiapas, Yucatán, Hidalgo, San Luis Potosí, Querétaro, Veracruz, Sinaloa, México, Michocán.
Región 3	Baja California, Quintana Roo.
Región 4	Distrito Federal, Oaxaca, Puebla, Tlaxcala, Guanajuato, Morelos, Nuevo León, Tabasco.

Al igual que en 4.1.1.3 (*encadenamiento promedio* con la norma euclideana), se obtuvieron cuatro regiones que prácticamente coinciden; asimismo regresan Baja California y Quintana Roo a una misma región. Pero ahora se unen primero las regiones 1 y 2.

4.1.1.6. Análisis de Conglomerados con *liga peso - promedio ponderado* (Media de grupos), tomando como medida de similaridad la matriz de correlaciones.

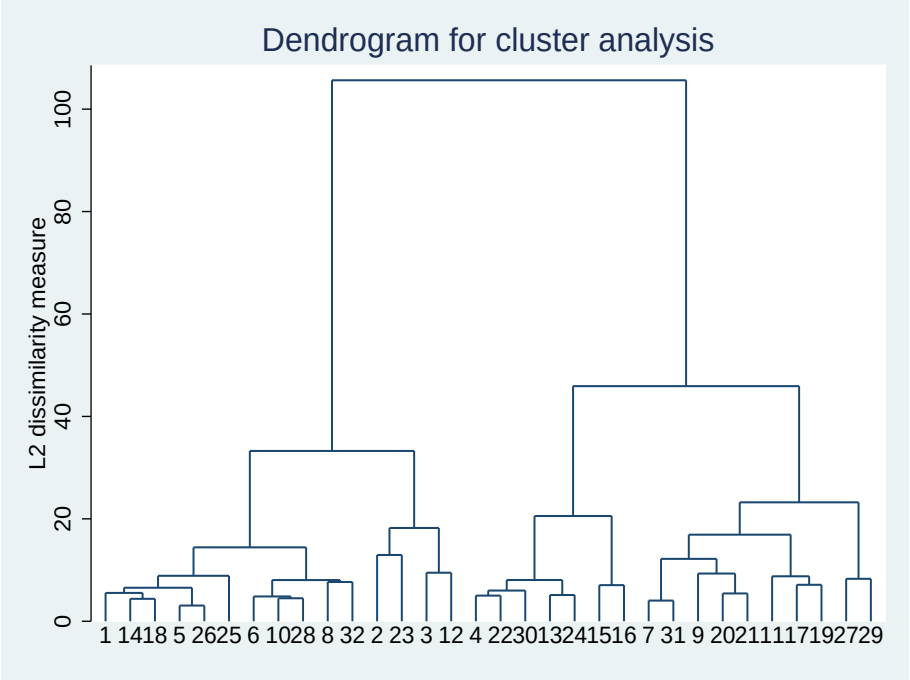


Al ejecutar el corte correspondiente respecto a la medida de similaridad, se obtienen las siguientes regiones:

Región 1	Aguascalientes, Sonora, Coahuila, Durango, Nayarit, Colima, Jalisco, Chihuahua, Tamaulipas, Guerrero, Zacatecas, Sinaloa, Baja California Sur.
Región 2	Baja California, Quintana Roo.
Región 3	Campeche, Querétaro, Chiapas, Yucatán, Hidalgo, San Luis Potosí, Veracruz.
Región 4	Distrito Federal, México, Guanajuato, Michocán, Morelos, Nuevo León.
Región 5	Oaxaca, Tlaxcala, Puebla, Tabasco.

Se obtienen las mismas regiones que en 4.1.1.4 (*encadenamiento promedio* con la matriz de correlaciones), pero ahora ya se unen de nuevo Baja California y Quintana Roo en una misma región.

4.1.1.7. Análisis de Conglomerados con el *Método de Ward*, tomando como medida de disimilaridad la norma euclideana.

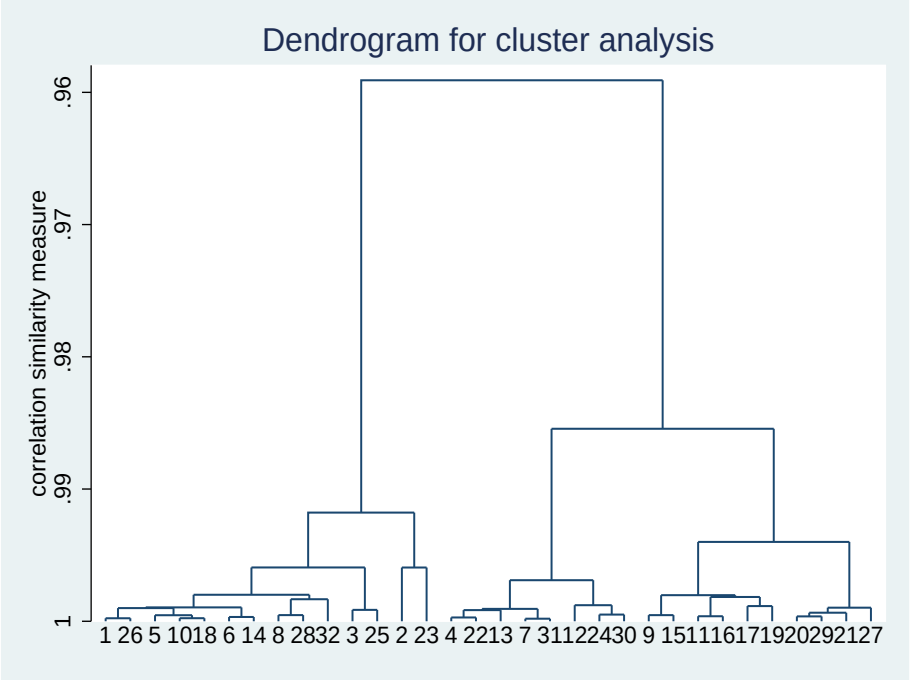


Realizando el corte correspondiente se obtienen las siguientes regiones:

Región 1	(Aguascalientes, Jalisco, Nayarit, Coahuila, Sonora, Sinaloa, Colima, Durango, Tamaulipas, Chiuahua, Zacatecas.
Región 2	Baja California, Quintana Roo.
Región 3	Baja California Sur, Guerrero).
Región 4	Campeche, Querétaro, Veracruz, Hidalgo, San Luis Potosí, México, Michoacán.
Región 5	Chiapas, Yucatán, Distrito Federal, Oaxaca, Puebla, Guanajuato, Morelos, Nuevo León.

Se observa que los estados de Baja California(2), Quintana Roo(23), Baja California Sur(3) y Guerrero(12) podrían unirse en una sola región. Incluso se pueden considerar tres regiones (como muestran los paréntesis) con base en la altura del corte al dendrograma.

4.1.1.8. Análisis de Conglomerados con el *Método de Ward*, tomando como medida de similaridad la matriz de correlaciones.



Al realizar el corte correspondiente respecto a la medida de similaridad se obtienen las siguientes regiones:

Región 1	(Aguascalientes, Sonora, Coahuila, Durango, Nayarit, Colima, Jalisco, Chihuahua, Tamaulipas, Zacatecas, Baja California Sur, Sinaloa.
Región 2	Baja California, Quintana Roo).
Región 3	(Campeche, Querétaro, Hidalgo, Chiapas, Yucatán, Guerrero, San Luis Potosí, Veracruz).
Región 4	(Distrito Federal, México, Guanajuato, Michoacán, Morelos, Nuevo León.
Región 5	Oaxaca, Tlaxcala, Puebla, Tabasco).

Nótese que se mantienen Baja California y Quintana Roo en una misma región, así como la consistencia de las regiones 1 y 3. Pero a diferencia de los casos previos, ahora es posible conformar dos (incluso tres) grandes regiones, como se muestra en el cuadro con los paréntesis.

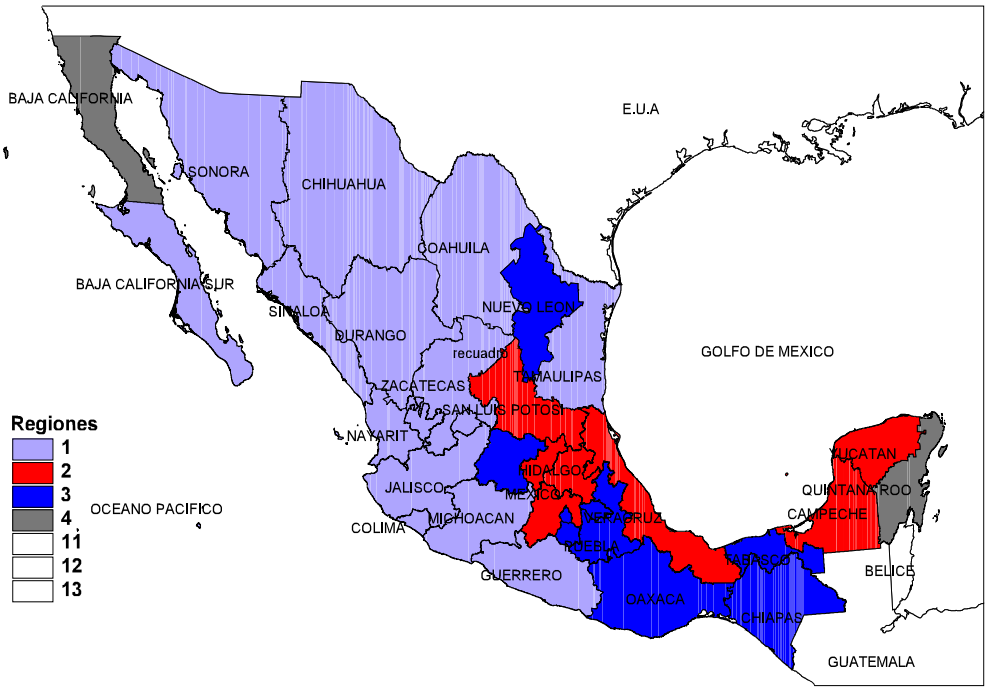
4.1.2. Regiones obtenidas con el Algoritmo de Partición K - medias.

Enseguida se presentan los resultados obtenidos con el algoritmo de partición k - medias para 4 y 5 regiones[4].

- **Tomando como medida de disimilaridad la norma euclideana y k observaciones aleatorias como centros iniciales de grupos.**

Caso I. k=4 regiones.

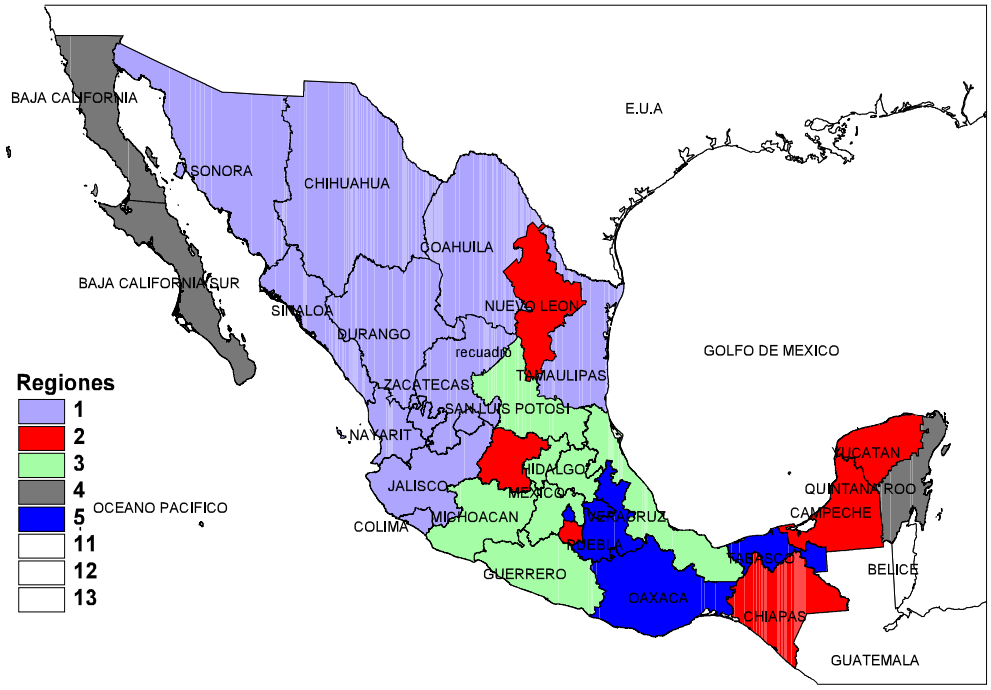
Región 1	Aguascalientes, Baja California Sur, Coahuila, Colima, Chihuahua, Durango, Guerrero, Jalisco, Michocán, Nayarit, Sinaloa, Sonora, Tamaulipas, Zacatecas.
Región 2	Campeche, Hidalgo, México, Querétaro, San Luis Potosí, Veracruz, Yucatán.
Región 3	Chiapas, Distrito Federal, Guanajuato, Morelos, Nuevo León, Oaxaca, Puebla, Tabasco, Tlaxcala.
Región 4	Baja California, Quintana Roo.



[4]Las regiones 11, 12 y 13 corresponden a Océano Pacífico, Golfo de México y E.U.A., respectivamente.

Caso II. k=5 regiones.

Región 1	Aguascalientes, Coahuila, Colima, Chihuahua, Durango, Jalisco, Nayarit, Sinaloa, Sonora, Tamaulipas, Zacatecas.
Región 2	Campeche, Chiapas, Guanajuato, Morelos, Nuevo León, Yucatán.
Región 3	Guerrero, Hidalgo, México, Michoacán, Querétaro, San Luis Potosí, Veracruz.
Región 4	Baja California, Quintana Roo, Baja California Sur.
Región 5	Distrito Federal, Oaxaca, Puebla, Tabasco, Tlaxcala.

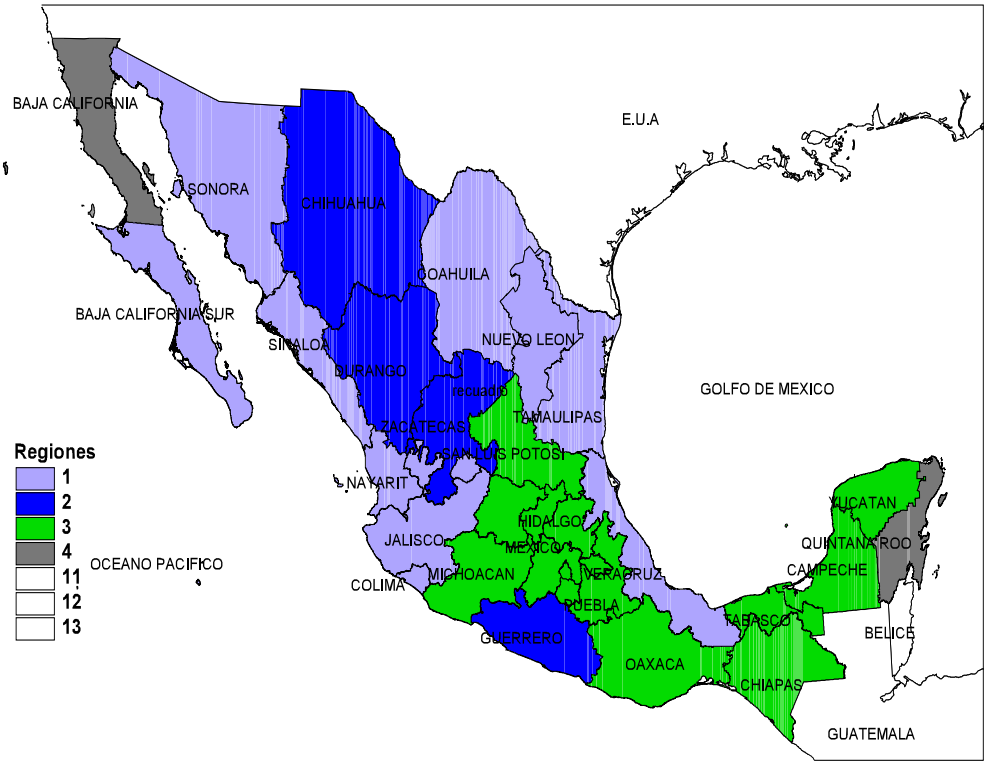


Al pasar de cuatro a cinco regiones los estados de Oaxaca, Puebla, Tabasco y Tlaxcala forman una sola región, lo cual apoya los resultados previos. La región 3 del **Caso I** se separa para formar parte de las regiones 2 y 5 del **Caso II**. Además la región 1 ($k = 4$) se segmenta para formar parte de las regiones 1 y 3 del caso $k = 5$. En ambos casos, los estados de Baja California y Quintana Roo mantienen su comportamiento al separarse del resto.

- **Tomando como medida de similaridad la matriz de correlaciones y k observaciones aleatorias como centros iniciales de grupos.**

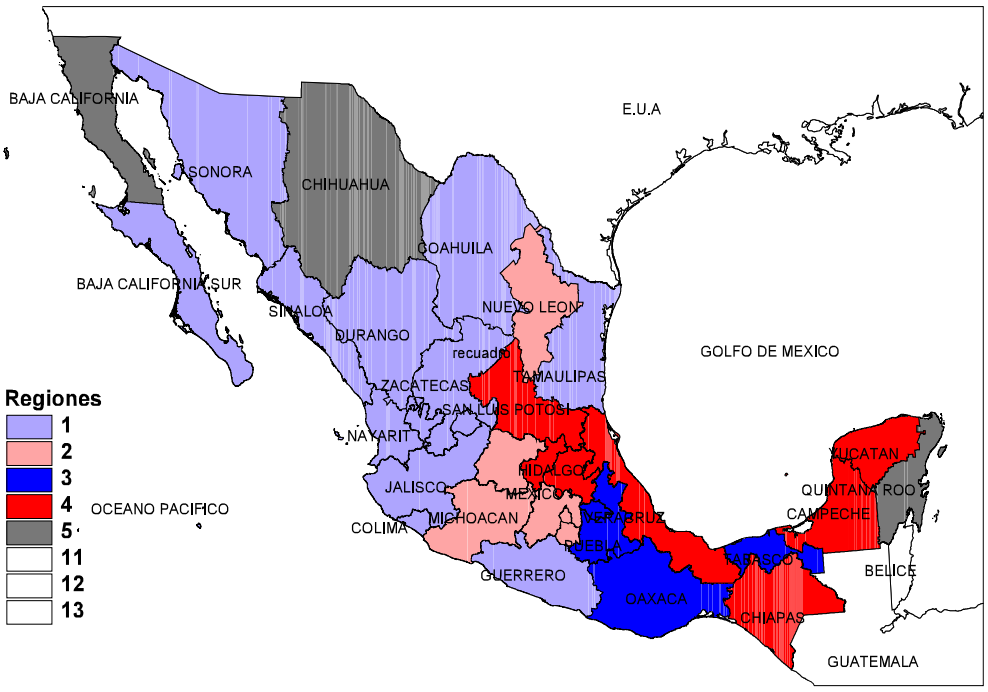
Caso I. k=4 regiones.

Región 1	Aguascalientes, Baja California Sur, Coahuila, Colima, Jalisco, Nayarit, Nuevo León, Sinaloa, Sonora, Tamaulipas, Veracruz.
Región 2	Chihuahua, Durango, Guerrero, Zacatecas.
Región 3	Campeche, Chiapas, Distrito Federal, Guanajuato, Hidalgo, México, Michoacán, Morelos, Oaxaca, Puebla, Querétaro, San Luis Potosí, Tabasco, Tlaxcala, Yucatán.
Región 4	Baja California, Quintana Roo.



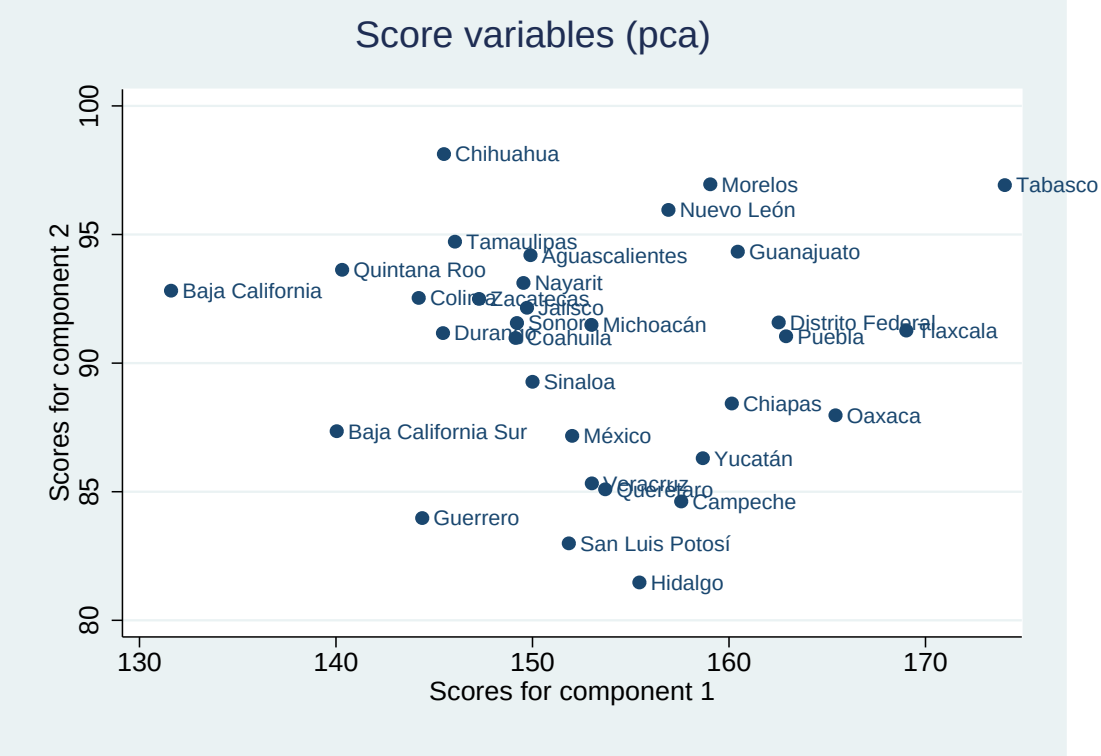
Caso II. k=5 regiones.

Región 1	Aguascalientes, Baja California Sur, Coahuila, Colima, Durango, Guerrero, Jalisco, Nayarit, Sinaloa, Sonora, Tamaulipas, Zacatecas.
Región 2	Distrito Federal, Guanajuato, México, Michoacán, Morelos, Nuevo León.
Región 3	Oaxaca, Puebla, Tabasco, Tlaxcala.
Región 4	Campeche, Chiapas, Hidalgo, Querétaro, San Luis Potosí, Veracruz, Yucatán.
Región 5	Baja California, Chihuahua, Quintana Roo.



Las regiones 1 y 2 del **Caso I** forman en esencia la región 1 para $k = 5$, salvo por el estado de Chihuahua, el cual se une a Baja California y Quintana Roo. Nótese también que la unión de las regiones 2, 3 y 4 del **Caso II**, dan como resultado la región 3 del **Caso I**. Nuevamente en ambos casos, los estados de Baja California y Quintana Roo mantienen su comportamiento al separarse del resto.

4.1.3. Gráfica de los dos primeros componentes principales obtenidos con la matriz de varianzas y covarianzas.



El porcentaje de varianza explicado por las primeras dos componentes es 82 %.

De la gráfica anterior se observa:

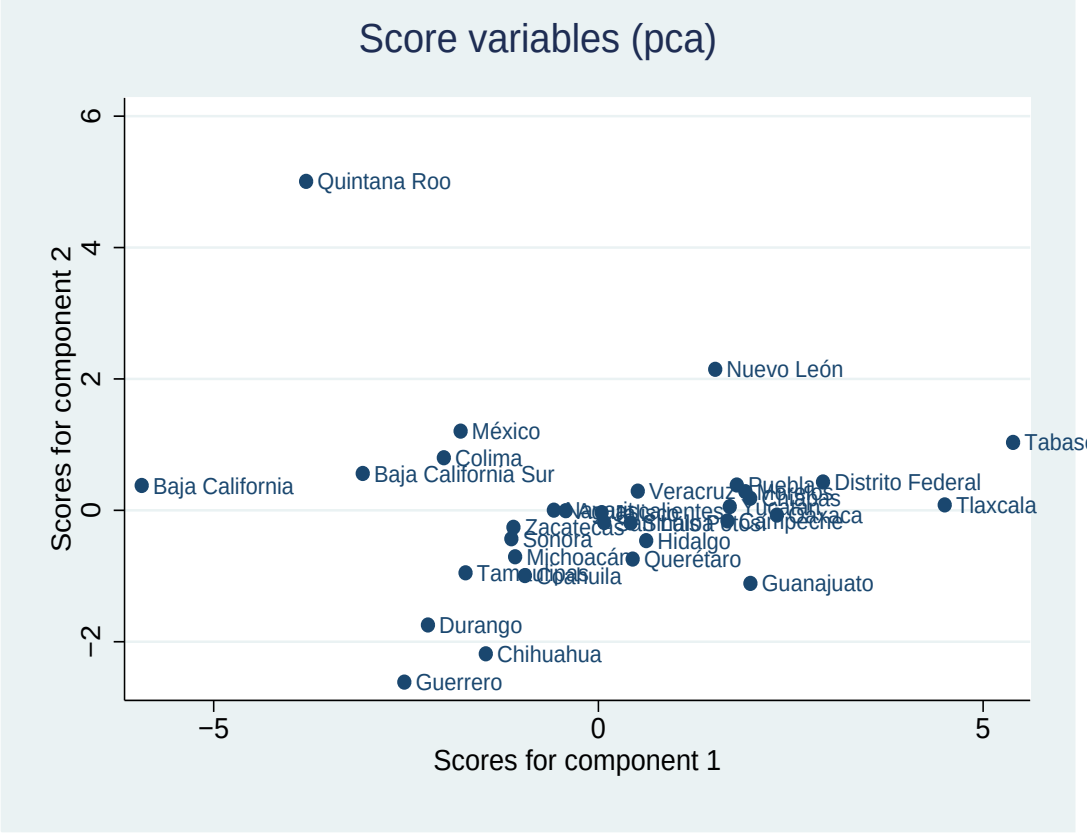
1. La cercanía que guardan Baja California y Quintana Roo, lo cual apoya la hipótesis de que ambos mantienen comportamientos similares.

2. En la zona central de la gráfica (incluyendo Chihuahua) se observan gran parte de los estados que conforman la **Región 1** en el Análisis de Conglomerados, tanto en métodos jerárquicos como con el algoritmo de las k-medias.

3. Asímismo, con base en los análisis previos, en la zona inferior derecha se observan los estados que, en esencia, formaron una región: Chiapas, Yucatán, Veracruz, Querétaro, San Luis Potosí, Campeche, e Hidalgo.

4. En la zona superior derecha se observa otro grupo de estados que prácticamente se asociaron para conformar una región: Distrito Federal, Morelos, Nuevo León, Guanajuato, Puebla, Tlaxcala y Tabasco.

4.1.4. Gráfica de los dos primeros componentes principales obtenidos con la matriz de correlaciones.



El porcentaje de varianza explicado por las primeras dos componentes es 63 %.

Nótese:

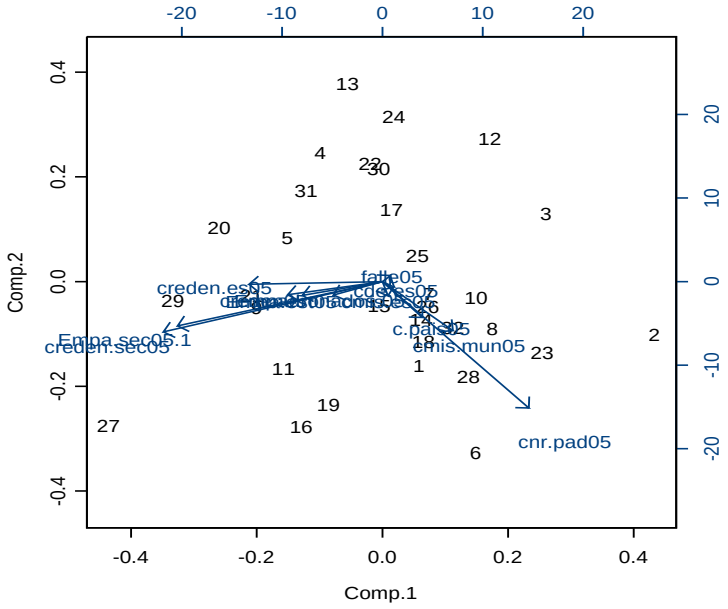
1. La consistencia en el comportamiento de Baja California y Quintana Roo, que a pesar de no estar tan cercanos en esta gráfica, se verifica su "*alejamiento*" del resto de los estados.

2. En la parte central de la gráfica, una concentración de los estados que conformaron, en la mayoría de los análisis, la **Región 1**.

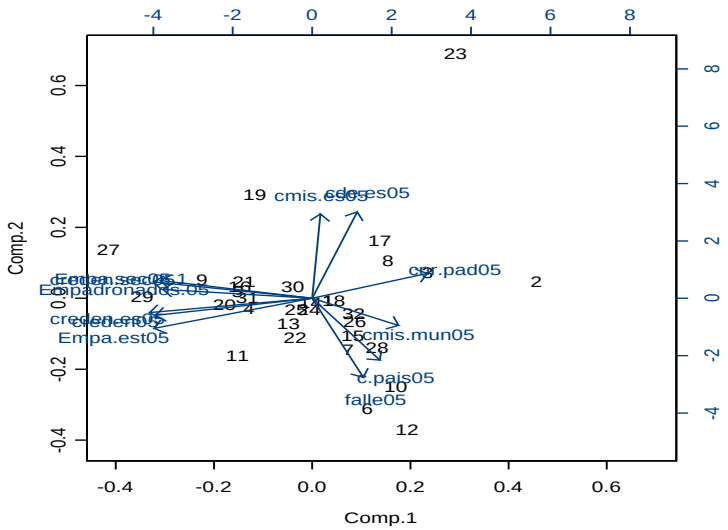
3. En ambas gráficas, el **primer componente** juega un papel importante en las discriminaciones.

4.1.5. Gráficas Biplot de los componentes principales obtenidos con la VNM2005.

4.1.5.1. Biplot con la matriz de varianzas y covarianzas.



4.1.5.2. Biplot con la matriz de correlaciones.



Las variables tienen la siguiente notación:

Notación	Variables (representadas con los vectores.)
Empadronados.05	Empadronados en 2005.
Empa.es05	Empadronados en el estado en 2005 .
Empa.sec05	Empadronados en la sección en 2005.
creden05	Credencializados en 2005.
creden.es05	Credencializados en el estado en 2005.
creden.sec05	Credencializados en la sección en 2005.
falle05	Fallecidos en 2005.
cnr.pad05	Cambios de domicilio no reportados en el padrón en 2005.
cmis.mu05	Cambios de domicilio al mismo municipio en 2005.
cmis.es05	Cambios de domicilio a otro municipio dentro del mismo estado en 2005.

Notación	**Variables (representadas con los vectores.)**
cde.es05	Cambios de domicilio a otro estado en 2005.
c.pais05	Cambios de domicilio a otro país en 2005.

En el Biplot con la matriz de covarianzas se observan dos grandes grupos de variables no correlacionados, donde los indicadores de mayor impacto son Cambios no reportados al Padrón, Credencializados y Empadronados. No se observa una fuerte asociación entre indicadores y estados. Mientras que en el Biplot obtenido con la matriz de correlaciones se aprecia con mayor énfasis el comportamiento de Baja California (2) y Quintana Roo (23), lo cual se ha detectado; asimismo se tienen grupos de variables no correlacionadas (en este caso son tres grupos) y se nota mayor asociación entre indicadores y entidades.

Con estos resultados (VNM2005), se verifica la tendencia a formar cinco regiones, y más aún, en cada de ellas se mantiene un número fijo de entidades federativas. Principalmente, se observa el "alejamiento" de los estados de Baja California y Quintana Roo así como el *constante* agrupamiento de los estados de Oaxaca, Tabasco, Tlaxcala y Puebla y los estados de México, Distrito Federal, Nuevo León, Guanajuato y Morelos. Las regiones que resaltan son:

1. Aguascalientes, Sonora, Coahuila, Durango, Nayarit, Colima, Jalisco, Chihuahua, Tamaulipas, Guerrero, Zacatecas, Sinaloa, Baja California Sur.
2. Baja California y Quintana Roo.
3. Campeche, Querétaro, Chiapas, Yucatán, Hidalgo, SLP y Veracruz.
4. Oaxaca, Tabasco, Tlaxcala y Puebla.
5. Distrito Federal, México, Guanajuato, Nuevo León y Morelos.

Las restantes entidades federativas mantuvieron ligeras variaciones de una región a otra durante el desarrollo de los análisis. Sin embargo, es posible comenzar a vislumbrar un *comportamiento* de las entidades federativas que, aunque no es definitivo, desde un inicio presenta diferencias significativas con la propuesta de regiones realizada por CONAPO [5], que es la siguiente:

Noroeste I:	Baja California, Baja California Sur y Sonora.
Noroeste II:	Sinaloa y Nayarit.
Norte:	Chihuahua, Coahuila, Nuevo León y Tamaulipas.
Norte - centro:	Durango, San Luis Potosí y Zacatecas.
Occidente:	Aguascalientes, Colima y Jalisco.
Centro:	Guanajuato, Michoacán y Querétaro.
Metropolitana:	Distrito Federal, México y Morelos.
Oriente:	Hidalgo, Puebla, Tlaxcala y Veracruz.
Sur:	Chiapas, Guerrero y Oaxaca.
Península:	Campeche, Quintana Roo, Tabasco y Yucatán.

En la siguiente sección veremos los resultados que arroja el análisis de los datos para la VNM2006. Las preguntas naturales que surgen tales como ¿Se mantendrán las regiones?, ¿ Se mantendrá el comportamiento de Baja California y Quintana Roo?, serán abordadas.

[5]CONAPO, Desigualdad Regional y Marginación Municipal en México, 1990. México, D.F. Noviembre de 1994.

4.2. Resultados obtenidos con la VNM2006.

Ahora se exhiben los resultados obtenidos con la VNM2006, en el siguiente orden:

1. Los dendrogramas correspondientes a los Análisis de Conglomerados así como las regiones formadas al realizar los cortes. Nuevamente, es posible construir un numero mayor o menor de clusters, lo cual queda en función de la altura del corte; sin embargo se optó por mantener las estructuras de las regiones que por realizar los cortes a una misma altura.

2. Las regiones obtenidas con el Algoritmo de Partición k-medias, para cinco y seis clusters.

3. Gráfica de los dos primeros componentes principales obtenidos con la matriz de varianzas y covarianzas.

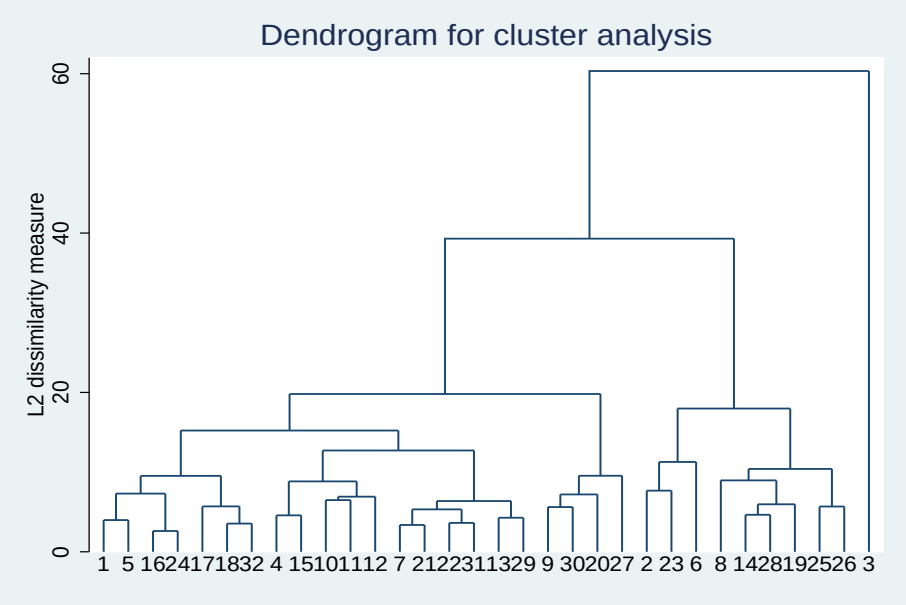
4. Gráfica de los dos primeros componentes principales obtenidos con la matriz de correlaciones.

5. Biplots suponiendo homoscedasticidad.

En este caso, el Algoritmo de las k-medias se "*corrió*" para generar cinco y seis clusters ya que se optó por mantener la estructura obtenida en los dendrogramas. En éstos, como se verá a continuación, el número de regiones obtenidas varió, precisamente, de cinco a seis.

4.2.1. Análisis de Conglomerados.

4.2.1.1. Análisis de Conglomerados con *liga completa*, tomando como medida de disimilaridad la norma euclideana.



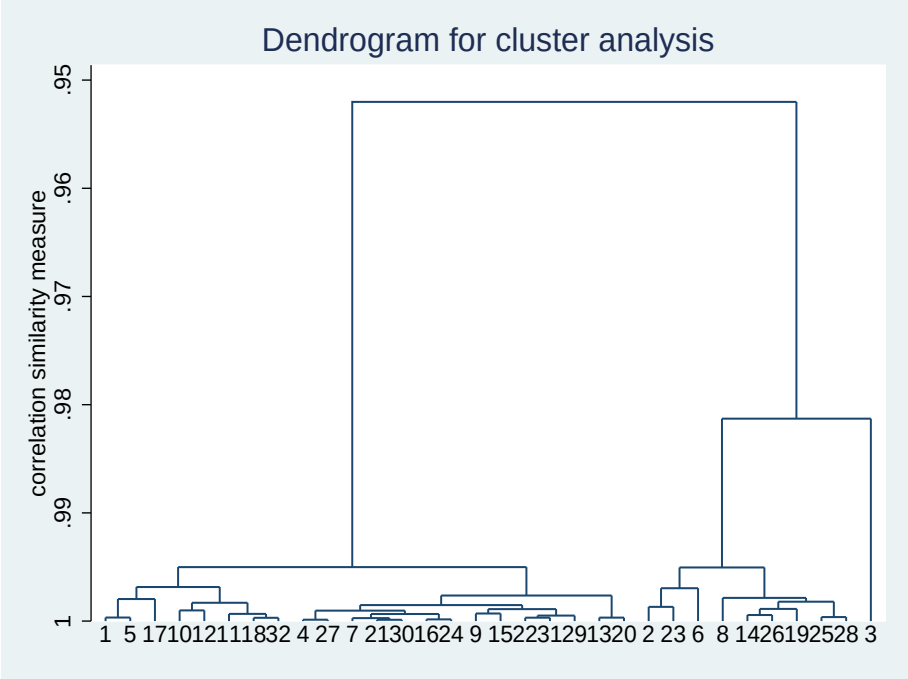
Realizando el corte correspondiente, se tienen las siguientes regiones:

Región 1	Aguascalientes, Coahuila, Michoacán, San Luis Potosí, Morelos, Zacatecas, Nayarit.
Región 2	Campeche, México, Durango, Guanajuato, Guerrero, Chiapas, Puebla, Querétaro, Yucatán, Hidalgo, Tlaxcala.
Región 3	Distrito Federal, Veracruz, Oaxaca, Tabasco.
Región 4	Baja California, Quintana Roo, Colima.
Región 5	Chihuahua, Jalisco, Tamaulipas, Nuevo León, Sinaloa, Sonora.
Región 6	Baja California Sur.

Nuevamente se observa la consistencia en la Región formada por Baja California y Quintana Roo [6], situación que se ha presentado desde los análisis con la VNM2005. Por otro lado, en la **Región 5** se observa una cohesión entre los estados del norte del país. La novedad es Baja California Sur, que ahora y a diferencia de la VNM2005, se separa del resto de los estados.

[6] A quienes ahora se les une Colima

4.2.1.2. Análisis de Conglomerados con *liga completa*, tomando como medida de similaridad la matriz de correlaciones.



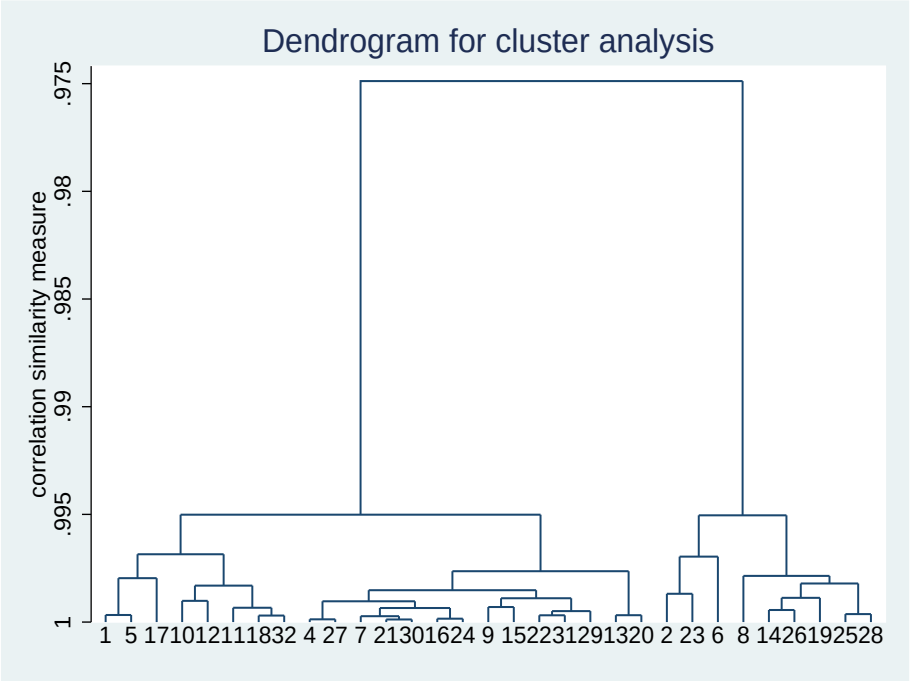
Al realizar el corte correspondiente respecto a la medida de similaridad, se obtienen las siguientes regiones:

Región 1	(Aguascalientes, Coahuila, Morelos, Durango, Guerrero, Guanajuato, Zacatecas, Nayarit.
Región 2	Campeche, Tabasco, Chiapas, Veracruz, Puebla, Michoacán, San Luis Potosí, Distrito Federal, México, Querétaro, Yucatán, Hidalgo, Tlaxcala, Oaxaca.)
Región 3	(Baja California, Quintana Roo, Colima.
Región 4	Chihuahua, Jalisco, Tamaulipas, Nuevo León, Sinaloa, Sonora.)
Región 5	Baja California Sur.

Nótese que las regiones 2 y 3 del caso anterior conforman ahora la **Región 2**. Se mantienen la cohesión de los estados del norte en la **Región 4**, el aislamiento

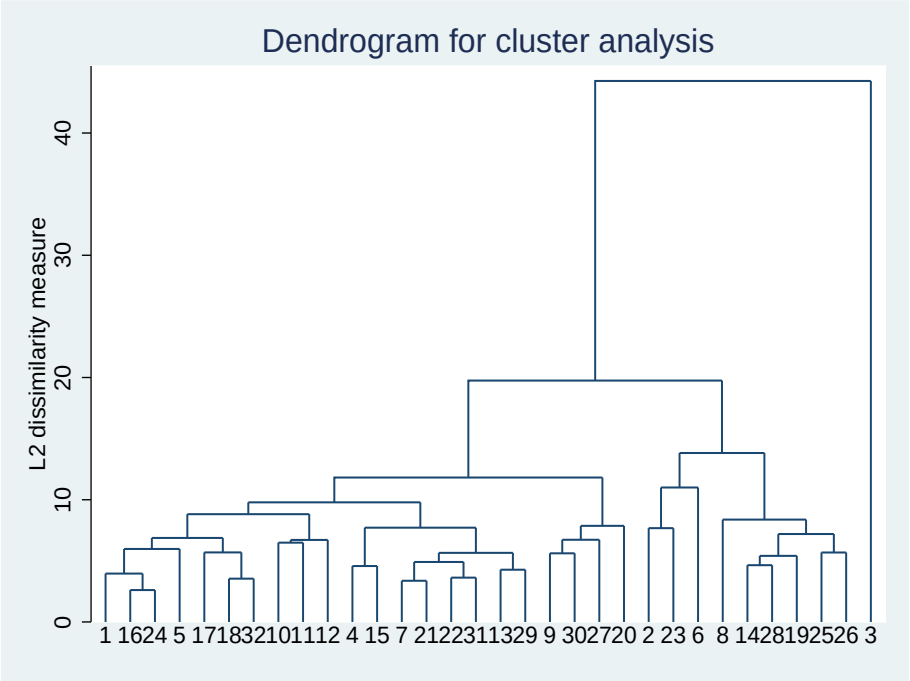
del estado de Baja California Sur y la **Región 3** conformada por Baja California, Quintana Roo y Colima.

Sin embargo, en el dendrograma no se observan con claridad estas regiones debido a que el estado de Baja California Sur (3) se "*une*" al finalizar el algoritmo. Por esta razón y con la intención de exhibir de manera más clara las regiones, se realizó una nueva "corrida" eliminando dicho estado. Se obtuvo el siguiente dendrograma:



Es claro que se mantienen las regiones ya mencionadas. Sin embargo, puede considerarse que las regiones 1 y 2 se mezclan para formar una sola, lo mismo ocurre con las regiones 3 y 4. Es decir, en este análisis puede hablarse de dos *grandes* regiones como lo muestran los paréntesis en la tabla anterior.

4.2.1.3. Análisis de Conglomerados con *liga promedio*, tomando como medida de disimilaridad la norma euclideana.

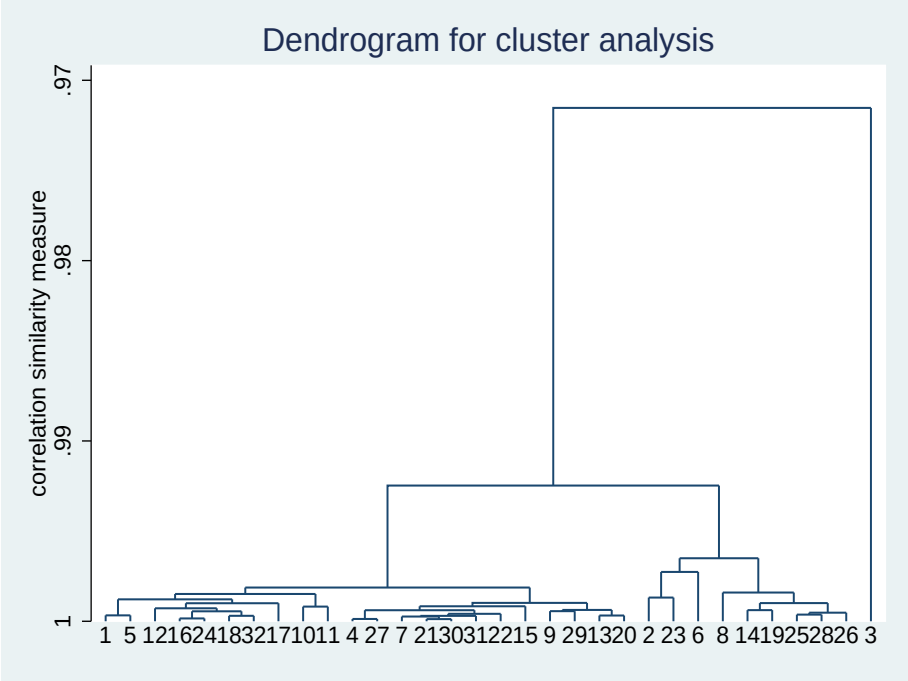


Tomando el corte correspondiente respecto a la medida de disimilaridad:

Región 1	(Aguascalientes, Michocán, San Luis Potosí, Coahuila, Morelos, Nayarit, Zacatecas, Durango, Guerrero, Guanajuato.
Región 2	Campeche, México, Chiapas, Puebla, Querétaro, Yucatán, Hidalgo, Tlaxcala.
Región 3	Distrito Federal, Veracruz, Tabasco, Oaxaca.)
Región 4	(Baja California, Quintana Roo, Colima.)
Región 5	(Chihuahua, Jalisco, Tamaulipas, Nuevo León, Sinaloa, Sonora.)
Región 6	(Baja California Sur.)

Las regiones 3, 4, 5 y 6 se mantienen, también, adviértase que la región 3 puede formar parte de la región 2 en función de la altura del corte. Pero más allá de mantener una estructura en los grupos, con base en el dendrograma es posible conformar cuatro regiones, como se muestra en el cuadro con los paréntesis.

4.2.1.4. Análisis de Conglomerados con *liga promedio*, tomando como medida de similaridad la matriz de correlaciones.



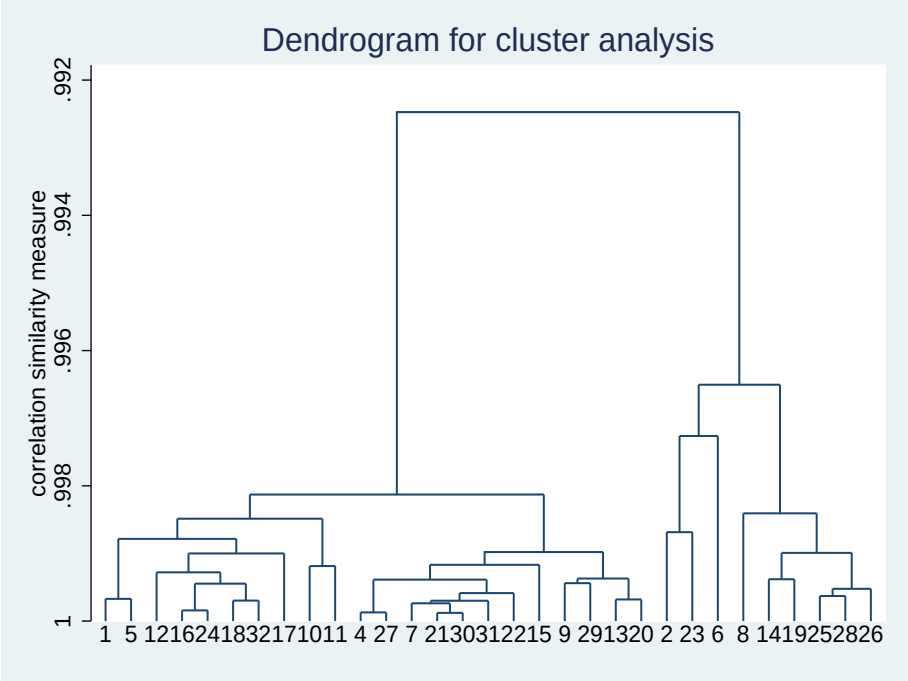
Al realizar el corte correspondiente, se tienen las siguientes regiones:

Región 1	(Aguascalientes, Michoacán, San Luis Potosí, Coahuila, Morelos, Nayarit, Zacatecas, Durango, Guerrero, Guanajuato.
Región 2	Campeche, México, Chiapas, Puebla, Querétaro, Yucatán, Veracruz, Tabasco, Distrito Federal, Tlaxcala, Hidalgo, Oaxaca.)
Región 3	(Baja California, Quintana Roo.
Región 4	Colima.
Región 5	Chihuahua, Jalisco, Tamaulipas, Nuevo León, Sinaloa, Sonora.)
Región 6	(Baja California Sur.)

Nótese que ahora Colima se separa ligeramente de Baja California y Quintana Roo, pero observando el dendrograma y por la cercanía entre éstos, puede asumirse sin perder consistencia, que dichos estados conforman una región. Se

presentan algunos cambios en la **Región 2**, mientras que 1, 5 y 6 se mantienen.

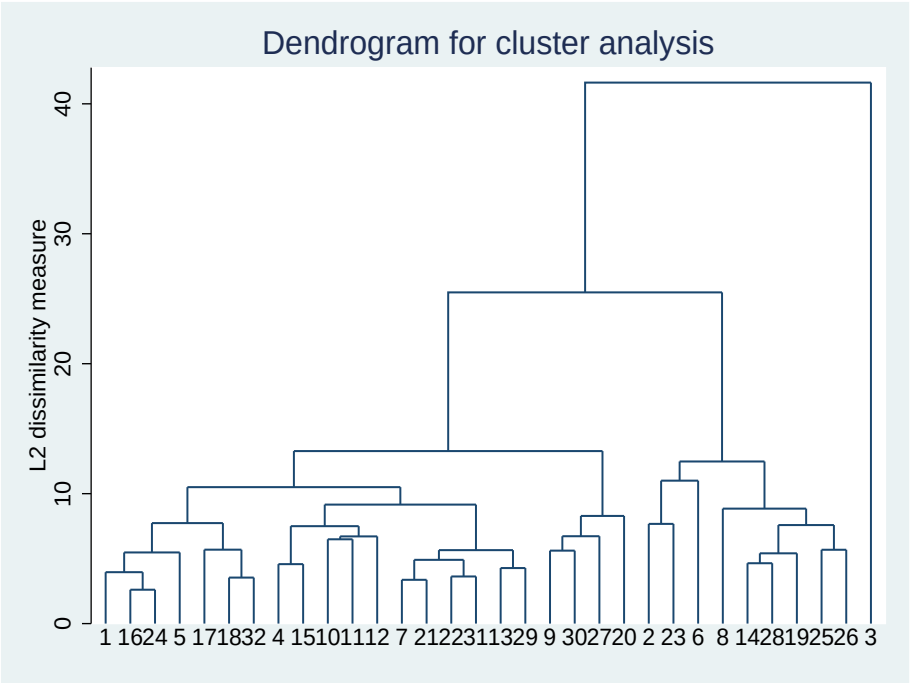
Sin embargo, el dendrograma no permite distinguir con claridad dichas regiones, y nuevamente se observa a Baja California Sur (3) como el estado que se "*une*" al final del proceso. En ese sentido y con la intención de exhibir de manera más clara las regiones, se omitió a dicho estado y se obtuvo el siguiente dendrograma:



Mejora notablemente la exhibición de las regiones, las cuales se conservan respecto a las que se obtuvieron sin omitir a Baja California Sur. Esto corrobora el "*aislamiento*" de dicho estado, y por ende, su influencia en estos análisis.

Finalmente, viendo el dendrograma es posible conformar incluso tres regiones como se muestra en la tabla anterior con los paréntesis. Como se mencionó, se ha optado por mantener la estructura en las regiones presentadas con los dendrogramas.

4.2.1.5. Análisis de Conglomerados con *liga peso - promedio ponderado* (Media de grupos), tomando como medida de disimilaridad la norma euclideana.

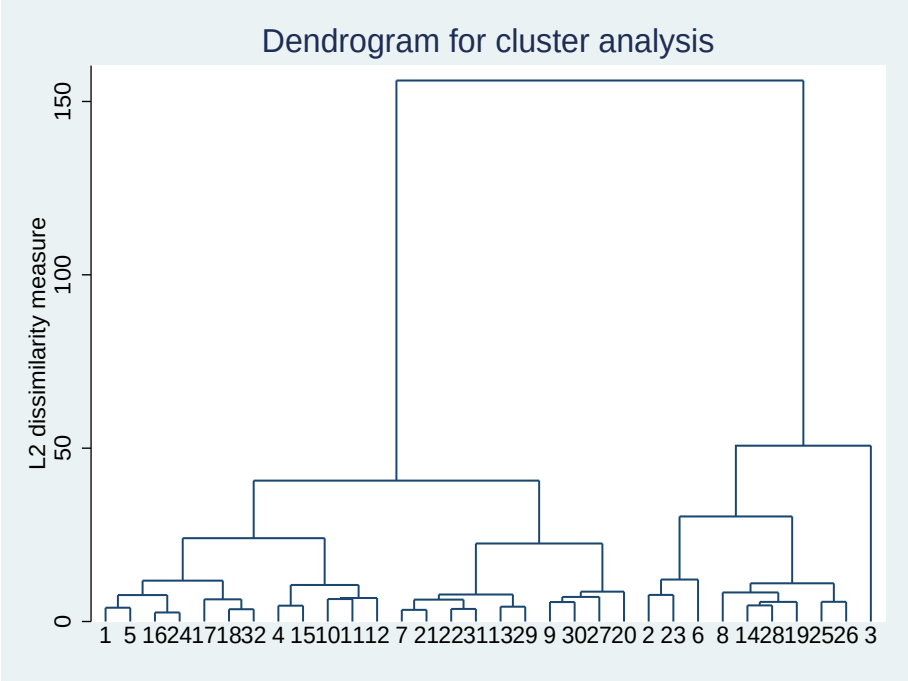


Tomando el corte respecto a la medida de disimilaridad se obtienen:

Región 1	Aguascalientes, Michoacán, San Luis Potosí, Coahuila, Morelos, Nayarit, Zacatecas.
Región 2	Durango, Guerrero, Guanajuato, Campeche, México Chiapas, Puebla, Querétaro, Yucatán, Hidalgo, Tlaxcala.
Región 3	Distrito Federal, Oaxaca, Tabasco, Veracruz.
Región 4	Baja California, Quintana Roo, Colima.
Región 5	Chihuahua, Jalisco, Tamaulipas, Nuevo León, Sinaloa, Sonora.
Región 6	Baja California Sur.

Las regiones 1, 3, 4, 5 y 6 mantienen su estructura respecto a análisis previos. Nuevamente se nota la inclusión de Colima en la región conformada por Baja California y Quintana Roo, así como la *separación* de BCS.

4.2.1.6. Análisis de Conglomerados con el *Método de Ward*, tomando como medida de disimilaridad la norma euclideana.

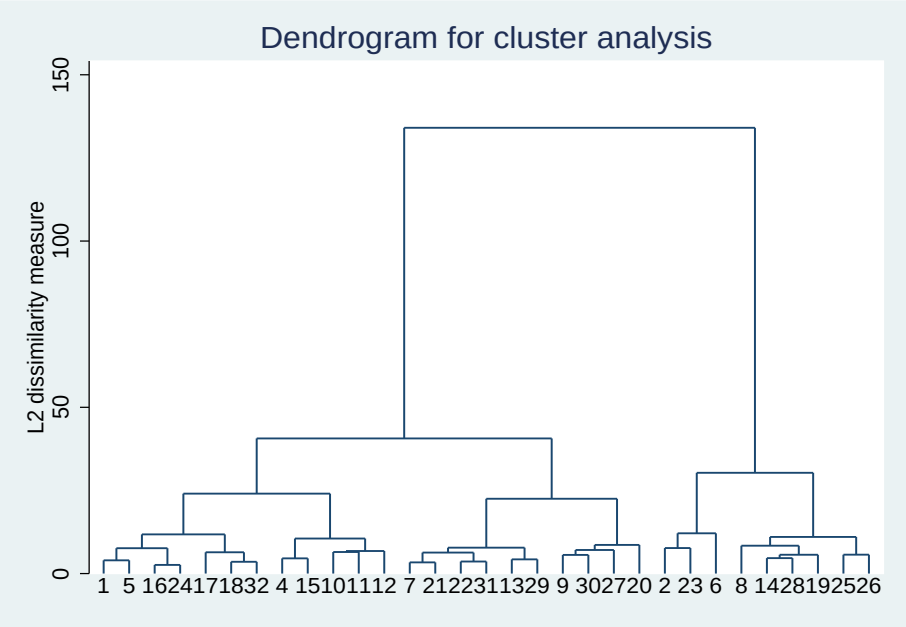


Tomando el corte respecto a la medida de disimilaridad se obtienen las siguientes regiones[7]:

Región 1	Aguascalientes, Michoacán, San Luis Potosí, Coahuila, Morelos, Nayarit, Zacatecas, Durango, Guerrero, Guanajuato, Campeche, México.
Región 2	Chiapas, Puebla, Querétaro, Yucatán, Hidalgo, Tlaxcala, Distrito Federal, Veracruz, Tabasco, Oaxaca.
Región 3	Baja California, Quintana Roo, Colima.
región 4	Chihuahua, Jalisco, Tamaulipas, Nuevo León, Sinaloa, Sonora.
Región 5	Baja California Sur.

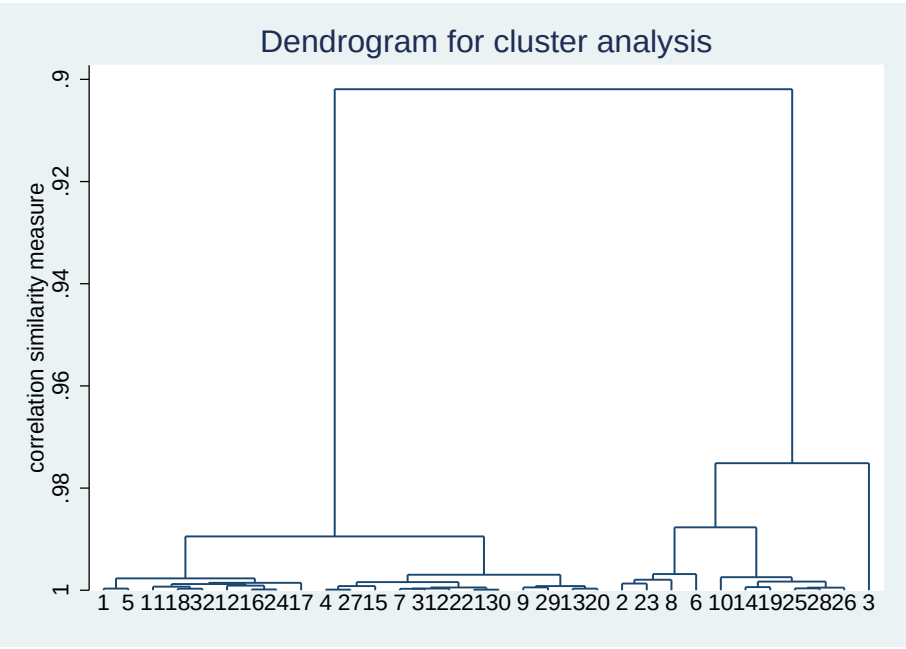
También en este caso se realizó una nueva "*corrida*" de Análisis de Conglomerados omitiendo a Baja California Sur (3), con la intención de corroborar el comportamiento de los estados en las regiones. El dendrograma es el siguiente:

[7]Resalta la consistencia de la región conformada por Colima, Baja California y Quintana Roo.



Como se esperaba, se conservan las regiones ya exhibidas. La exclusión de Baja California Sur no incide en el dendrograma.

4.2.1.7. Análisis de Conglomerados con el *Método de Ward*, tomando como medida de similaridad la matriz de correlaciones.

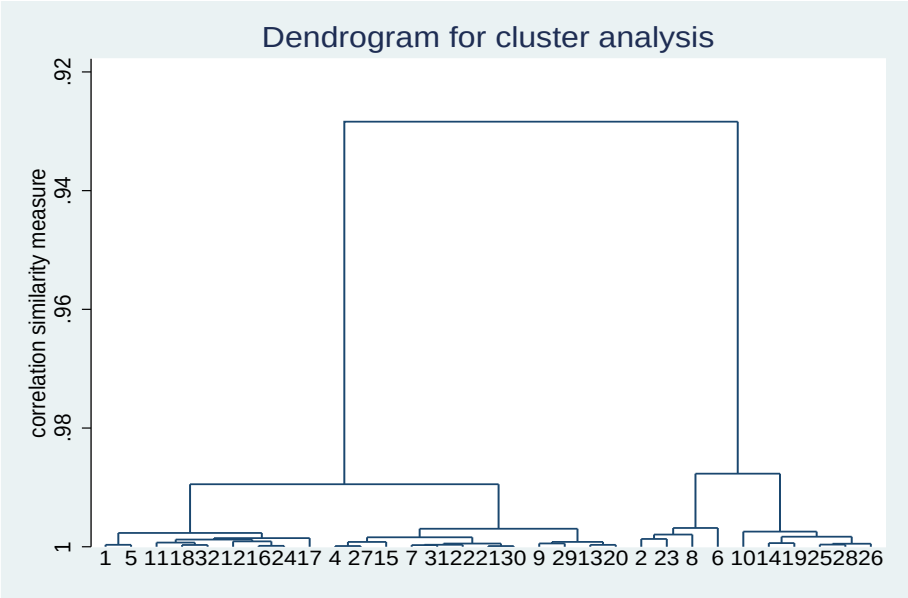


Al realizar el corte correspondiente resultan las siguientes regiones:

Región 1	Aguascalientes, Coahuila, Guanajuato, Nayarit, Zacatecas, Guerrero, Michoacán, San Luis Potosí, Morelos.
Región 2	Campeche, Tabasco, México, Chiapas, Yucatán, Querétaro, Puebla, Veracruz.
Región 3	Distrito Federal, Tlaxcala, Hidalgo, Oaxaca.
Región 4	Baja California, Quintana Roo, Chihuahua, Colima.
Región 5	Durango, Jalisco, Nuevo León, Sinaloa, Tamaulipas, Sonora.
Región 6	Baja California Sur.

Se confirman el comportamiento de BCS que se aleja del resto de los estados, la cohesión de Baja California, Quintana Roo y Colima (**Región 4**) y la consistencia de la **Región 6**. Y salvo ligeros cambios se mantienen las regiones 1 y 2.

Sin embargo, en este caso (a diferencia del caso anterior) *a priori* se observa que BCS (3) influye en la realización del dendrograma, por lo cual se realizó un nuevo análisis omitiendo a dicho estado, y aunque se mantienen las regiones, el dendrograma no mejoró significativamente.



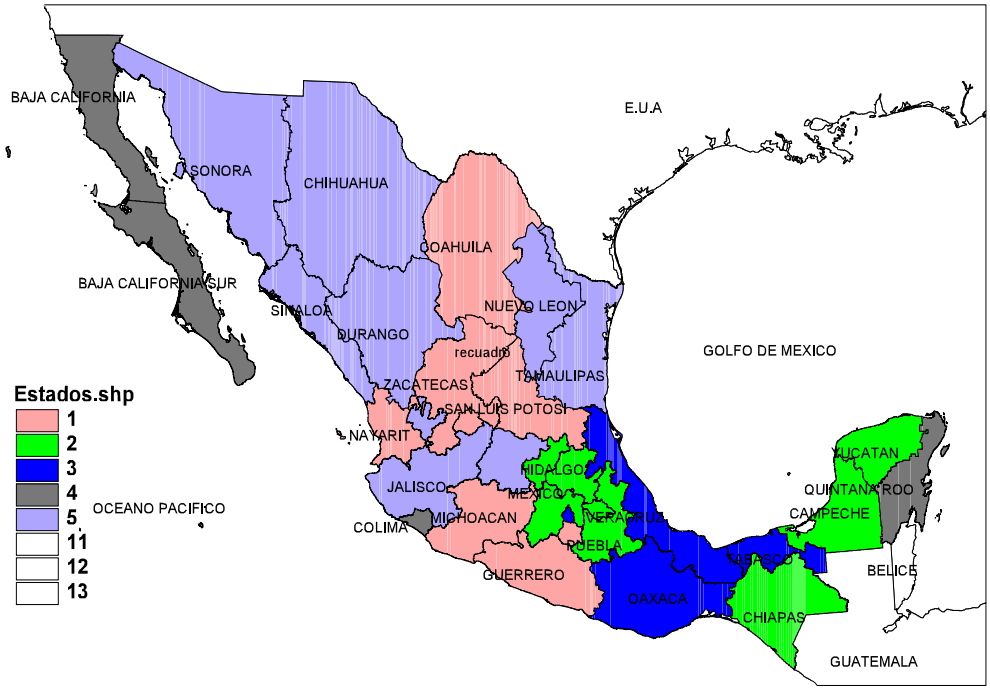
4.2.2. Regiones obtenidas con el Algoritmo de Partición K - medias.

A continuación se exhiben las regiones obtenidas con el algoritmo de partición clásico de las k-medias para 5 y 6 regiones.

- **Tomando como medida de disimilaridad la norma euclideana y k observaciones aleatorias como centros iniciales de grupos.**

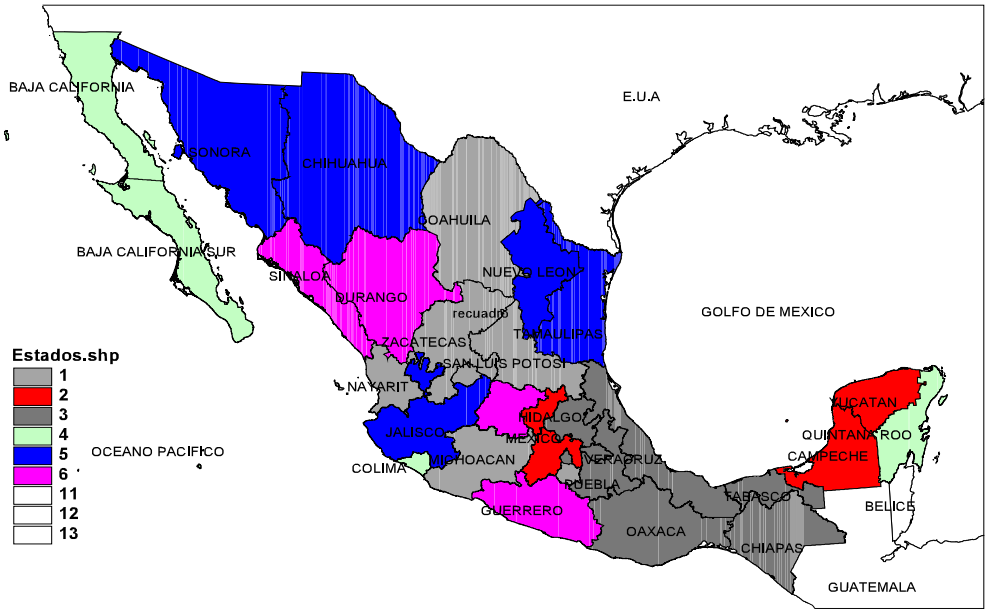
Caso I. k=5 regiones.

Región 1	Aguascalientes, Coahuila, Guerrero, Michoacán, Morelos, Nayarit, Zacatecas, San Luis Potosí.
Región 2	Campeche, Chiapas, Hidalgo, México, Puebla, Querétaro, Tlaxcala, Yucatán.
Región 3	Distrito Federal, Oaxaca, Tabasco, Veracruz.
Región 4	Baja California, Baja California Sur, Colima, Quintana Roo.
Región 5	Chihuahua, Durango, Guanajuato, Jalisco, Nuevo León, Sinaloa, Sonora, Tamaulipas.



Caso II. k=6 regiones.

Región 1	Aguascalientes, Coahuila, Michoacán, Morelos, Nayarit, Zacatecas, San Luis Potosí.
Región 2	Campeche, México, Querétaro, Yucatán.
Región 3	Distrito Federal, Hidalgo, Oaxaca, Puebla, Chiapas, Tabasco, Tlaxcala, Veracruz.
Región 4	Baja California, Baja California Sur, Colima, Quintana Roo.
Región 5	Chihuahua, Jalisco, Nuevo León, Sonora, Tamaulipas.
región 6	Durango, Guanajuato, Guerrero, Sinaloa.



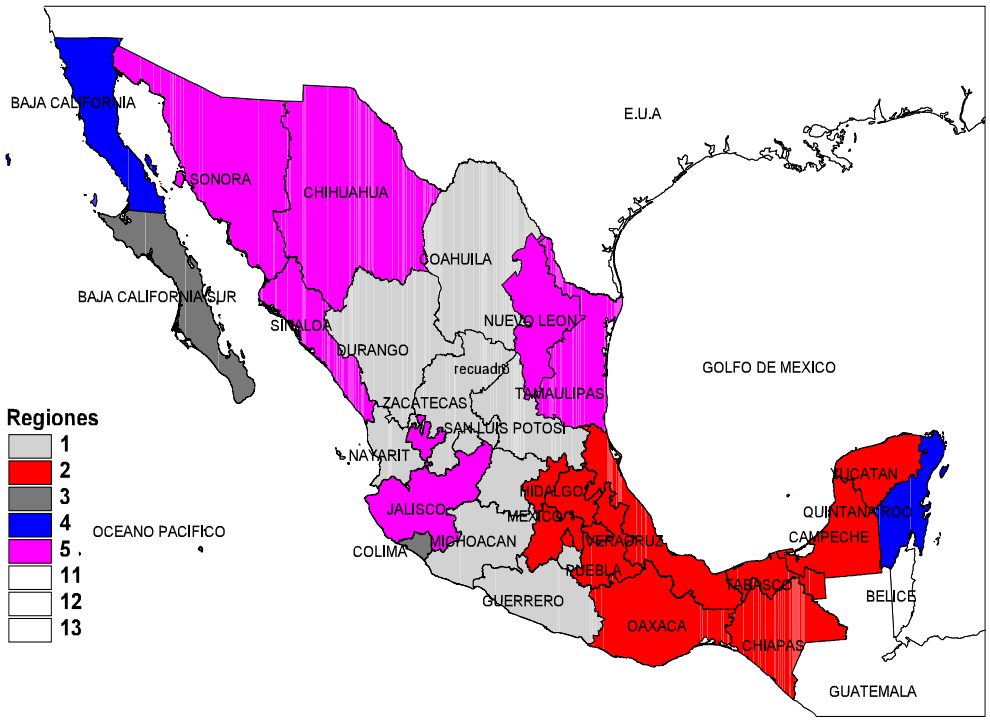
Observaciones:

1. La **Región 4** coincide en ambos casos. A excepción del estado de Guerrero, la **Región 1** también coincide en ambos casos.
2. En este análisis, resalta el hecho de que Baja California Sur se une a la región conformada por los estados de BC, Colima y Quintana Roo, la cual se mantuvo en la mayor parte de los resultados. Asimismo, la **Región 3** se mantiene en esencia (DF, Veracruz, Oaxaca y Tabasco).

- **Tomando como medida de similaridad la matriz de correlaciones y k observaciones aleatorias como centros iniciales de grupos.**

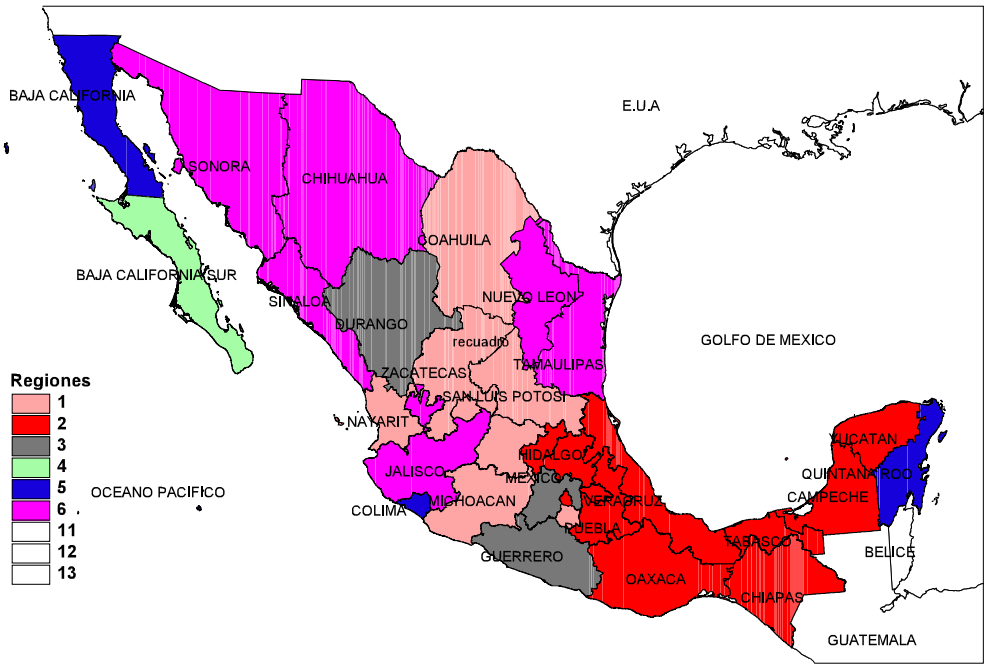
Caso I. k=5 regiones.

Región 1	Aguascalientes, Coahuila, Durango, Guanajuato, Guerrero, Michoacán, Morelos, Nayarit, San Luis Potosí, Zacatecas.
Región 2	Campeche, Chiapas, Distrito Federal, Hidalgo, México, Oaxaca, Puebla, Querétaro, Tabasco, Tlaxcala, Veracruz, Yucatán.
Región 3	Baja California Sur, Colima.
Región 4	Baja California, Quintana Roo.
Región 5	Chihuahua, Jalisco, Nuevo León, Sinaloa, Sonora, Tamaulipas.



Caso II. k=6 regiones.

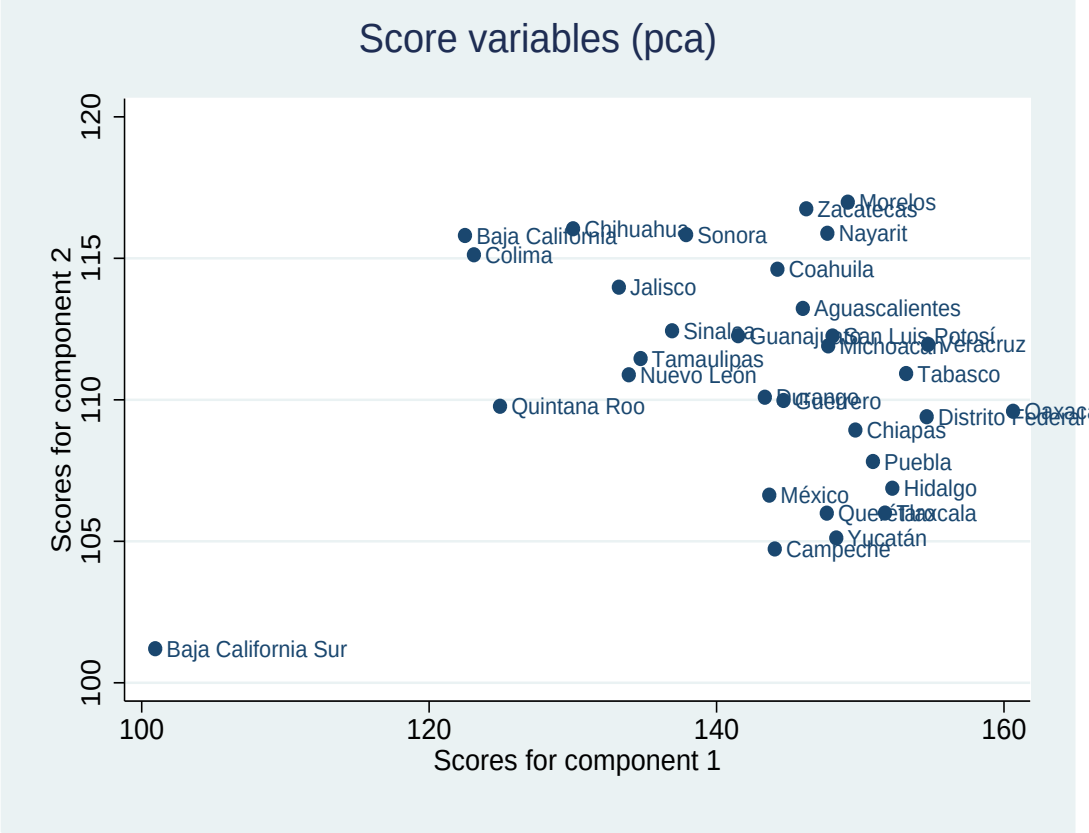
Región 1	Aguascalientes, Coahuila, Guanajuato, Michoacán, Morelos, Nayarit, San Luis Potosí, Zacatecas.
Región 2	Campeche, Chiapas, Distrito Federal, Hidalgo, Oaxaca, Puebla, Querétaro, Tabasco, Tlaxcala, Veracruz, Yucatán.
Región 3	Durango, Guerrero, México.
Región 4	Baja California Sur.
Región 5	Baja California, Colima, Quintana Roo.
Región 6	Chihuahua, Jalisco, Nuevo León, Sinaloa, Sonora, Tamaulipas.



Observaciones:

1. La **Región 6**, coincide con la **Región 5** del caso anterior. Las regiones 1 y 2 coinciden, salvo por los estados de Durango, Guerrero y México; los cuales, pasan a formar una región en el caso II.
2. Se corrobora el comportamiento similar de Baja California y Quintana Roo, así como el de Baja California Sur.

4.2.3. Gráfica de los dos primeros componentes principales obtenidos con la matriz de varianzas y covarianzas.



El porcentaje de varianza explicado por las primeras dos componentes es 90 %.

En la gráfica se observa:

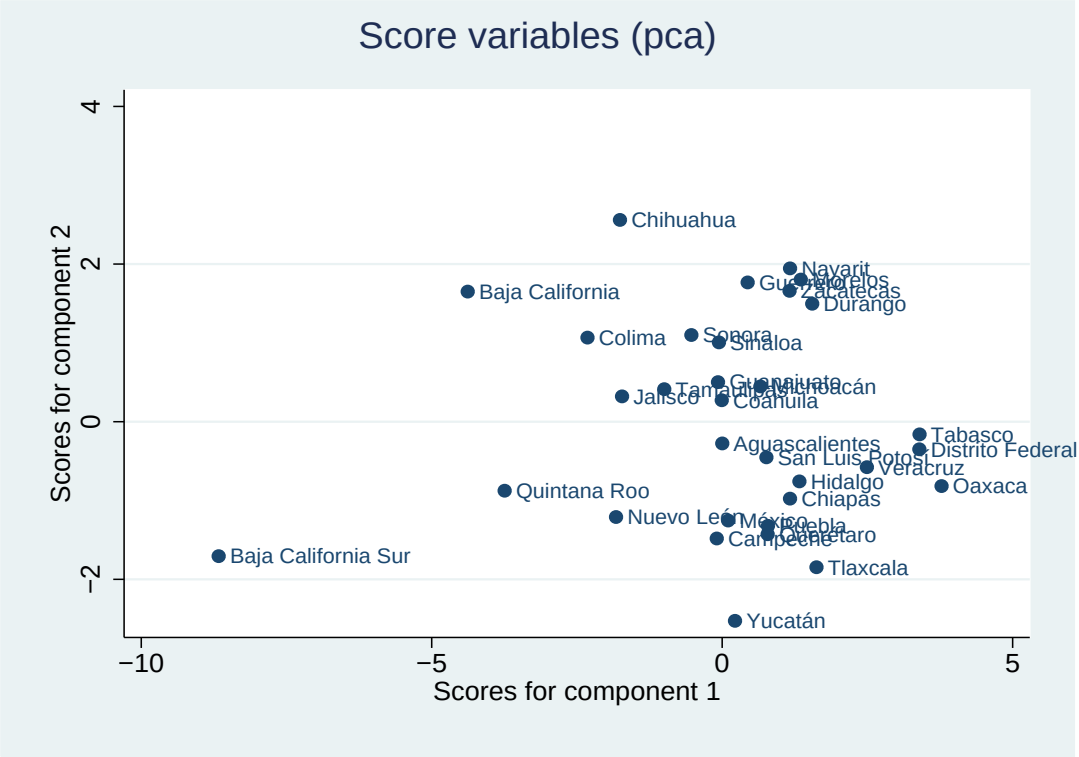
1. La importancia que cobran la primera y segunda componentes principales en el comportamiento de Baja California Sur, que se aleja notablemente del resto de los estados.

2. Asimismo, se nota la cohesión de Baja California, Colima y Quintana Roo.

3. En el centro de la gráfica se observa un grupo de estados que mantuvieron su comportamiento al pertenecer a la misma región, a decir: Chihuahua, Sonora, Jalisco, Sinaloa, Tamaulipas y Nuevo León.

4. En la parte inferior - derecha se observa otro grupo de estados que se mantuvo en una misma región: Chiapas, Puebla, Hidalgo, Tlaxcala, México, Querétaro y Yucatán.

4.2.4. Gráfica de los dos primeros componentes principales obtenidos con la matriz de correlaciones.

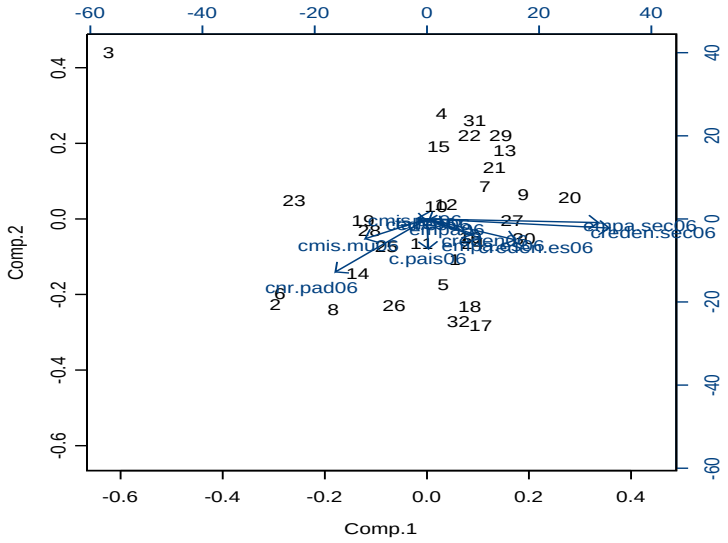


El porcentaje de varianza explicado por las primeras dos componentes es 65 %.

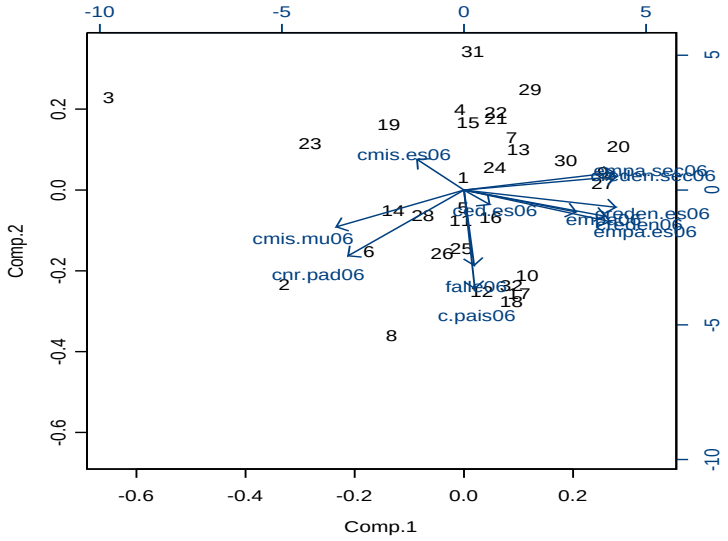
Nuevamente se observa la importancia del primer componente para determinar el "*alejamiento*" de Baja California Sur. En la parte central de ambas gráficas, se observa un grupo de estados que principalmente, reflejan el resultado obtenido en las regiones 1 y 2 del análisis de conglomerados. Se corrobora la cohesión de los estados de Sinaloa, Sonora, Nuevo León, Tamaulipas y Jalisco, que en la mayor parte de los análisis pertenecen a la misma región.

4.2.5. Gráficas Biplot de los componentes principales obtenidos con la VNM2006.

4.2.5.1. Biplot con la matriz de varianzas y covarianzas.



4.2.5.2. Biplot con la matriz de correlaciones.



Los nombres de las variables tienen la siguiente notación:

Notación	**Variables (representadas con los vectores.)**
empa06	Empadronados en 2006.
empa.es06	Empadronados en el estado en 2006 .
empa.sec06	Empadronados en la sección en 2006.
creden06	Credencializados en 2006.
creden.es06	Credencializados en el estado en 2006.
creden.sec06	Credencializados en la sección en 2006.
falle06	Fallecidos en 2006.
cnr.pad06	Cambios de domicilio no reportados en el padrón en 2006.
cmis.mu06	Cambios de domicilio al mismo municipio en 2006.
cmis.es06	Cambios de domicilio a otro municipio dentro del mismo estado en 2006.
ced.es06	Cambios de domicilio a otro estado en 2006.
c.pais06	Cambios de domicilio a otro país en 2006.

Respecto a los Biplots, se obtienen resultados similares que con la VNM2005, en particular cuando se emplea la matriz de varianzas y covarianzas con los indicadores *Cambios de domicilio a otro país (2005 y 2006), y Cambios de domicilio no reportados en el padrón (2005 y 2006)*. Cuando se emplea la matriz de correlaciones nuevamente se observan *conjuntos* de indicadores correlacionados. Se corrobora el *comportamiento* del estado de Baja California Sur (3), el cual se presentó en prácticamente en todos los análisis con la VNM2006 [8].

Ahora, comparando todos los resultados de esta sección (VNM2006) con los obtenidos para la VNM2005, hay primeramente un cambio en cuanto el número de regiones obtenidas, pues "*pasó*" de cinco a seis[9]. Sin embargo se tienen dos observaciones: por un lado, la diferencia se debe al estado de Baja California Sur que, respecto a la VNM2005, ahora se presenta aislado en prácticamente toda la

[8] También se observa un *alejamiento* de los estados de Quintana Roo (23), Colima (6) y Baja California (2).
[9] Nótese que la *sexta región* es la conformada por el estado de Baja California Sur.

sección, lo cual puede interpretarse como cinco regiones y un estado "*aislado*" que no pertenece a alguna. Por otro lado, ahora surge un región conformada exclusivamente por estados del norte: Chihuahua, Jalisco, Tamaulipas, Nuevo León, Sinaloa y Sonora. Las restantes entidades se mantienen en esencia, salvo ligeros cambios de una región a otra.

De esta manera, podemos exhibir y hablar de los grupos de estados que mostraron consistencia en este análisis; y en gran parte con el anterior(VNM2005):

1. Aguascalientes, Coahuila, Nayarit, Zacatecas, Guerrero, Durango, San Luis Potosí, Morelos y Guanajuato.
2. Campeche, Chiapas, Querétaro, Yucatán, Veracruz, Tabasco, Distrito Federal, Puebla y Oaxaca.
3. Baja California, Quintana Roo y Colima.
4. Chihuahua, Jalisco, Tamaulipas, Nuevo León, Sinaloa y Sonora.
5. Baja California Sur.

Nuevamente son claras las diferencia con las regiones arrojadas por CONAPO (presentadas en la sección anterior), principalmente, porque en ambas Verificaciones se ha detectado un *aislamiento* por parte de Baja California y Quintana Roo. Con estos análisis ya es posible especular que hablando en términos electorales, las entidades federativas se comportan de manera completamente diferente a cuando se habla en términos de marginación.

4.3. Resultados obtenidos con los indicadores ponderados

En esta sección se presentan los resultados obtenidos con los indicadores ponderados por el recíproco de su error estándar[10]. Con esto se pretende que los análisis se realicen *más a modo* con los datos[11], en el sentido de darle un mayor "*peso*" a aquellos indicadores con menor error estándar, pues se sabe que un estimador con un menor error estándar (es decir, con menor desviación) tendrá una mayor probabilidad de producir una estimación más cercana al parámetro de población que se está considerando[12] .

Para ambos casos, VNM2005 y VNM2006, la secuencia de exhibición de gráficas y resultados seguirá la tónica presentada en las secciones previas: Análisis de Conglomerados, Algoritmo de Partición Clásico de k-medias, gráficas de los dos primeros componentes principales y Biplots.

4.3.1. Resultados con los indicadores ponderados para VNM2005

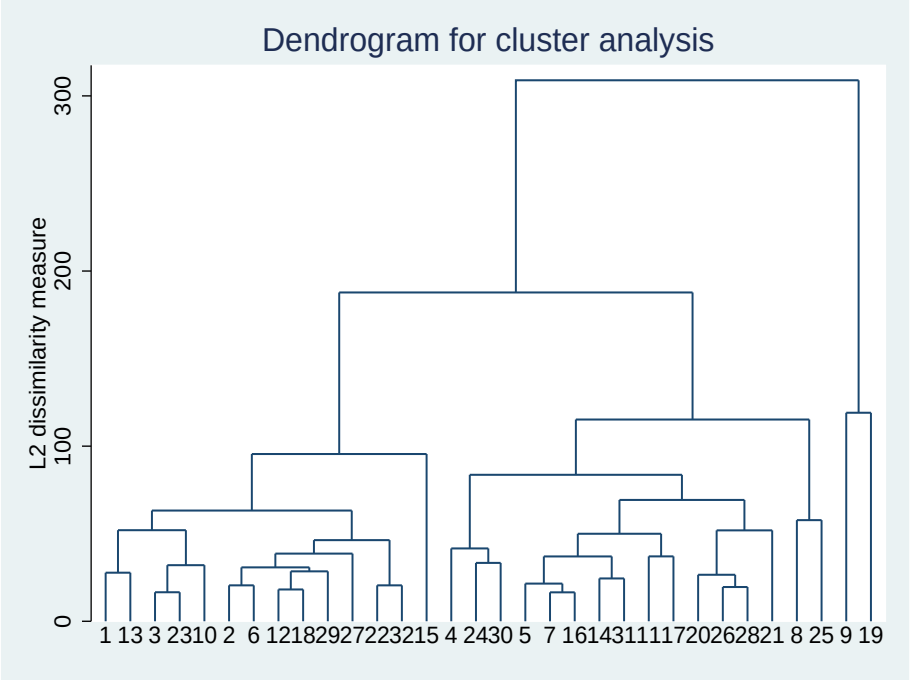
Se ha mencionado, pero vale la pena recordar que en los resultados de los Análisis de Conglomerados que se exhibirán a continuación (tanto para VNM2005 como para VNM2006), se optó por mantener la misma estructura de las regiones que por realizar los cortes en los dendrogramas a una misma altura. Un razonamiento análogo se siguió al prefijar el número de regiones considerado en el Algoritmo de las k-medias.

[10] Los errores estándar por indicador se incluyeron en los resultados proporcionados por IFE. "Verificación Nacional Muestral 2005", Informe final de resultados, 30 de agosto de 2005 y "Verificación Nacional Muestral 2006", Informe de resultados, 2 de mayo de 2006.

[11] Cabe mencionar que las diferencias y/o semenjanzas en los resultados de esta sección pueden también adjudicarse a una distorsión del fenómeno debido a aspectos muestrales, y no necesariamente tenga relación directa con los errores estándar.

[12] Concepto comúnmente conocido como ***Eficiencia*** del estimador.

4.3.1.1. Análisis de Conglomerados con *liga completa*, tomando como medida de disimilaridad la norma euclideana.

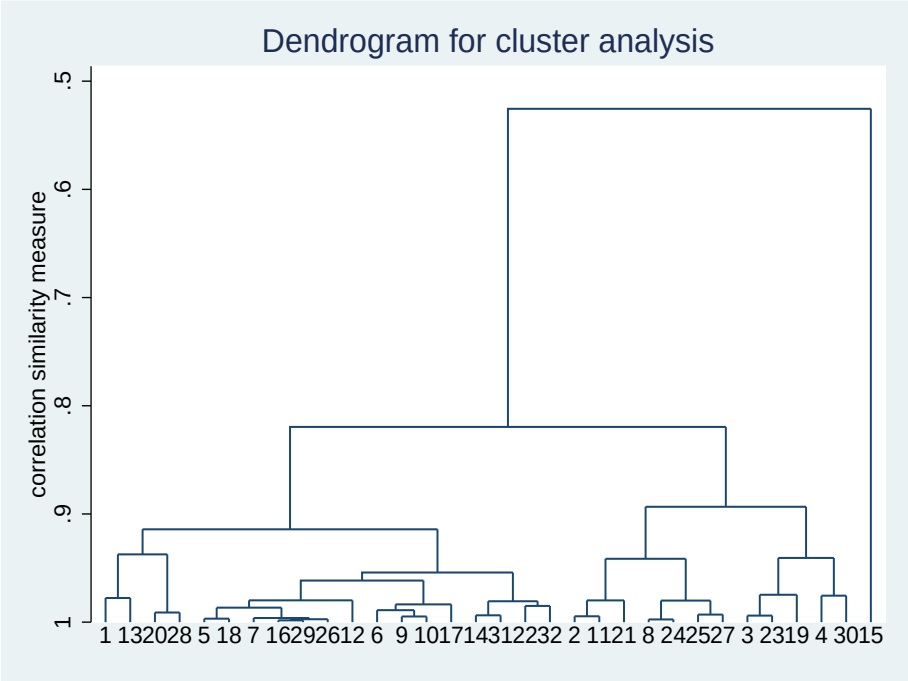


Realizando el corte correspondiente se obtienen las siguientes regiones:

Región 1	(Aguascalientes, Hidalgo, Baja California Sur, Quintana Roo, Durango, Baja California, Colima, Guerrero, Nayarit, Tlaxcala, Tabasco, Querétaro, Zacatecas, Oaxaca.
Región 2	México).
Región 3	(Campeche, San Luis Potosí, Veracruz, Coahuila, Chiapas, Michocán, Jalisco, Yucatán, Guanajuato, Morelos, Sonora, Tamaulipas, Puebla.
Región 4	Chihuahua, Sinaloa).
Región 5	Distrito Federal, Nuevo León.

Se nota desde un principio una discrepancia con los resultados obtenidos con los indicadores NO ponderados. El estado de México podría colocarse en la región 1 aunque esta situación depende de la altura del corte. Lo mismo se puede decir sobre las regiones 3 y 4, las cuales podrían formar una sola región.

4.3.1.2. Análisis de Conglomerados con *liga completa*, tomando como medida de similaridad la matriz de correlaciones.

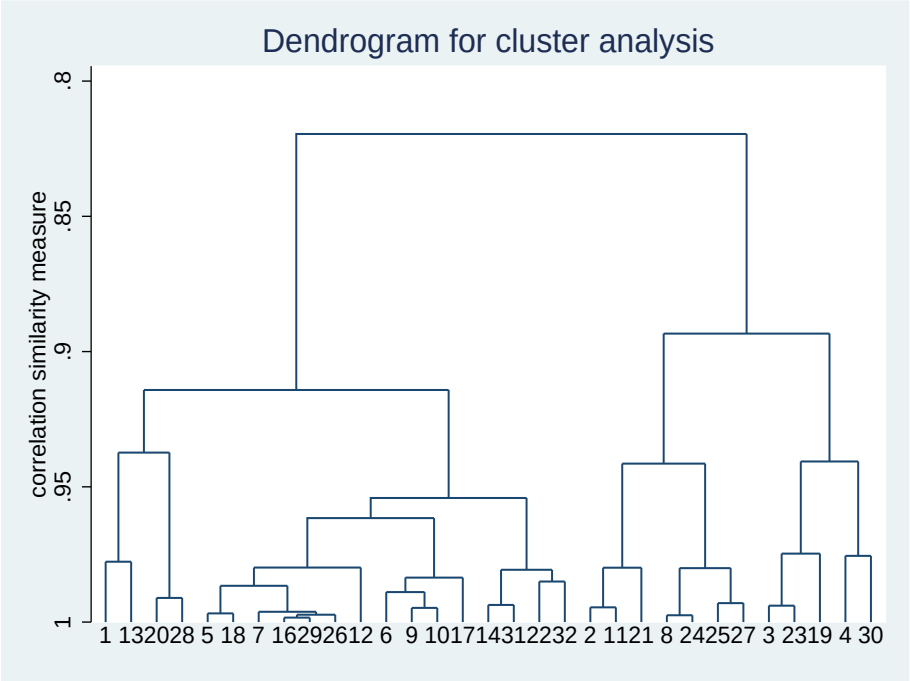


Al realizar el corte correspondiente respecto a la medida de similaridad, se obtienen las siguientes regiones:

Región 1	Aguascalientes, Hidalgo, Oaxaca, Tamaulipas.
Región 2	Coahuila, Nayarit, Chiapas, Michoacán, Tlaxcala, Sonora, Guerrero, Colima, Distrito Federal, Durango, Morelos, Jalisco, Yucatán, Querétaro, Zacatecas.
Región 3	Baja California, Guanajuato, Puebla, Chihuahua, San Luis Potosí, Sinaloa, Tabasco.
Región 4	Baja California Sur, Quintana Roo, Nuevo León, Campeche, Veracurz.
Región 5	México.

Este análisis apoya el comportamiento aislado del estado de México, el cual fue apenas detectado en el caso anterior. Tan es así, que puede pensarse que pertenece a la región 1.

Es claro que el estado de México (15) modifica sensiblemente el dendrograma. Para tratar de cuantificar esta modificación, se realizó nuevamente el mismo análsis eliminando el estado de México. Se obtuvo el siguiente dendrograma:



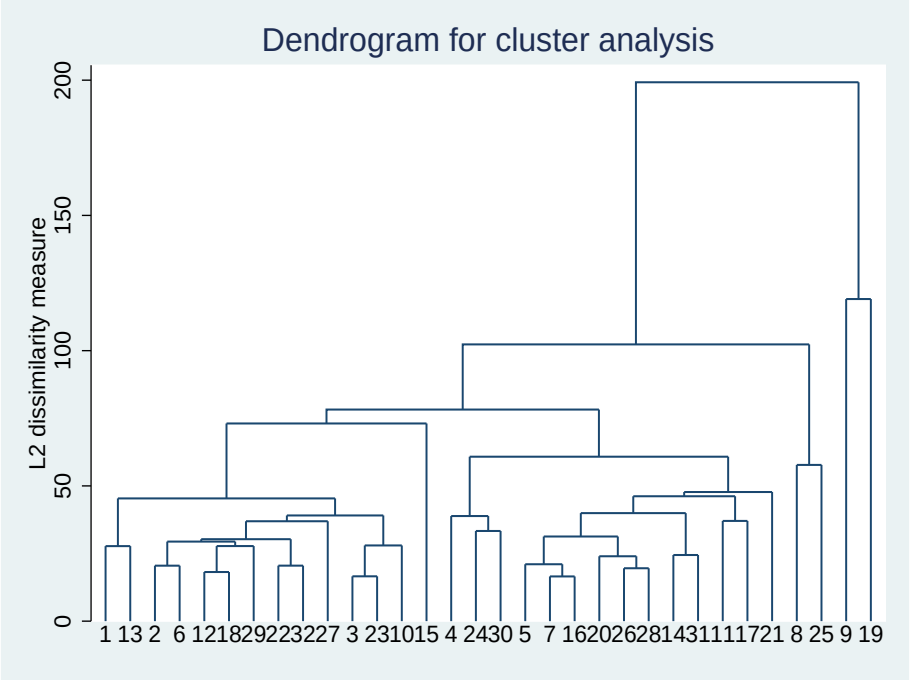
Y las siguientes regiones[13]:

Región 1	Aguascalientes, Hidalgo, Oaxaca, Tamaulipas.
Región 2	Coahuila, Nayarit, Chiapas, Michoacán .
Región 3	Tlaxcala, Sonora, Guerrero, Colima, Distrito Federal, Durango, Morelos, Jalisco, Yucatán, Queréataro, Zacatecas.
Región 4	Baja California, Guanajuato, Puebla, Chihuahua, Tabasco, Sinaloa, San Luis Potosí, Baja California Sur, Quintana Roo, Nuevo León, Campeche, Veracruz.

Se podría formar un mayor número de regiones bajando el punto de corte en el dendrograma, pero como se verá en los próximos resultados, no se percibe un comportamiento que permita argumentar y justificar el formar un mayor número de regiones y con ello, romper con la estructura de regiones que se ha mantenido.

[13]Se forman cuatro grupos de estados que se comportan al interior de manera similar (o dos grandes grupos).

4.3.1.3. Análisis de Conglomerados con *liga promedio*, tomando como medida de disimilaridad la norma euclideana.

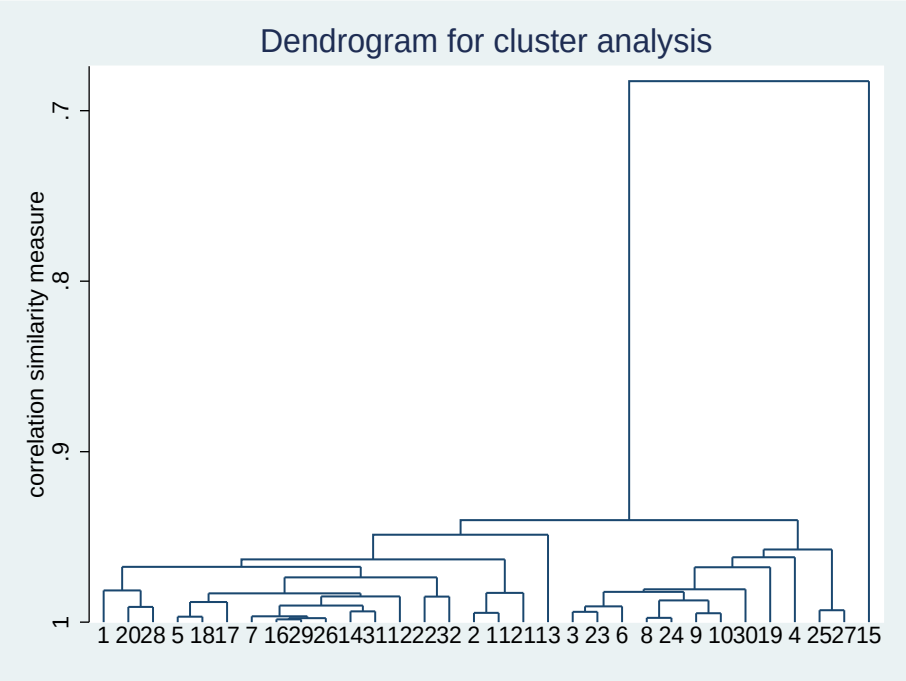


Al ejecutar el corte correspondiente se obtienen las siguientes regiones:

Región 1	(Aguascalientes, Hidalgo, Baja California, Colima, Guerrero, Nayarit, Tlaxcala, Querétaro, Zacatecas, Tabasco, Quintana Roo, Durango, Baja California Sur.
Región 2	México.)
Región 3	Campeche, San Luis Potosí, Veracruz, Coahuila, Chiapas, Michocán, Jalisco, Yucatán, Guanajuato, Morelos, Sonora, Tamaulipas, Puebla, Oaxaca.
Región 4	Chihuahua, Sinaloa.
Región 5	Distrito Federal, Nuevo León.

Se presenta un efecto de *encadenamiento* y respecto al análisis realizado en 4.3.1.1 (con la *liga completa*, norma euclideana) se conservan prácticamente todas las regiones. El comportamiento del estado de México se preserva, aunque podría vérsele como parte de la región 1, como muestran los paréntesis.

4.3.1.4. Análisis de Conglomerados con *liga promedio*, tomando como medida de similaridad la matriz de correlaciones.

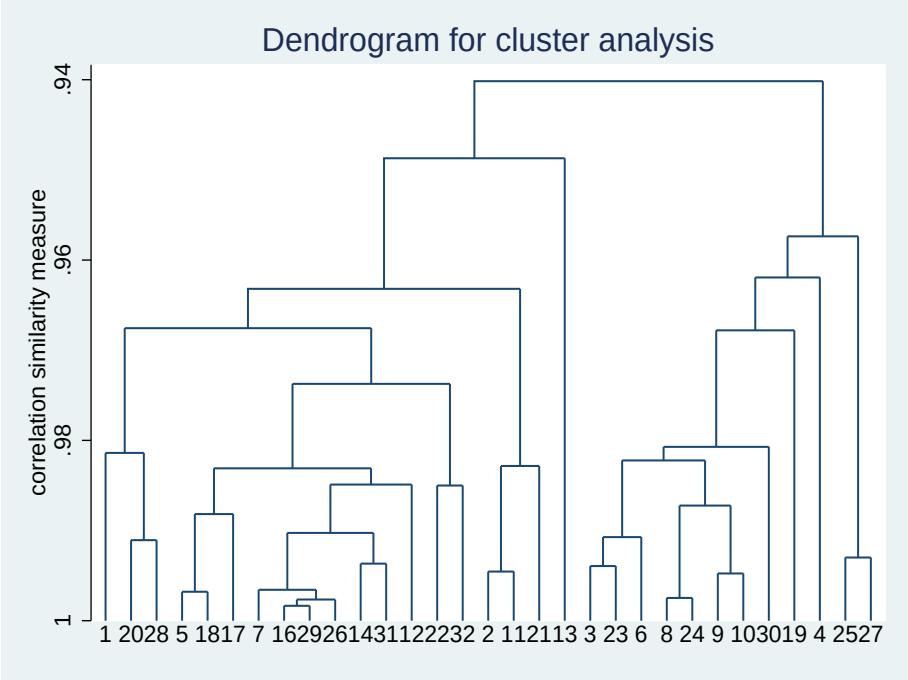


Con el corte correspondiente, se obtienen las siguientes regiones:

Región 1	Aguascalientes, Oaxaca, Tamaulipas, Coahuila, Morelos, Nayarit, Chiapas, Michoacán, Tlaxcala, Sonora, Jalisco, Yucatán, Guerrero, Querétaro, Zacatecas, Baja California, Guanajuato, Puebla, Hidalgo.
Región 2	Baja California Sur, Quintana Roo, Colima, Chihuahua, San Luis Potosí, Distrito Federal, Durango, Veracruz, Nuevo León, Campeche, Sinaloa, Tabasco.
Región 3	México.

A diferencia de resultados en las secciones previas, ahora se obtuvieron dos regiones claramente visibles y el estado de México, el cual se separa de ambas. Apenas puede verse una relación con las regiones obtenidas en 4.3.1.1. y 4.3.1.2.

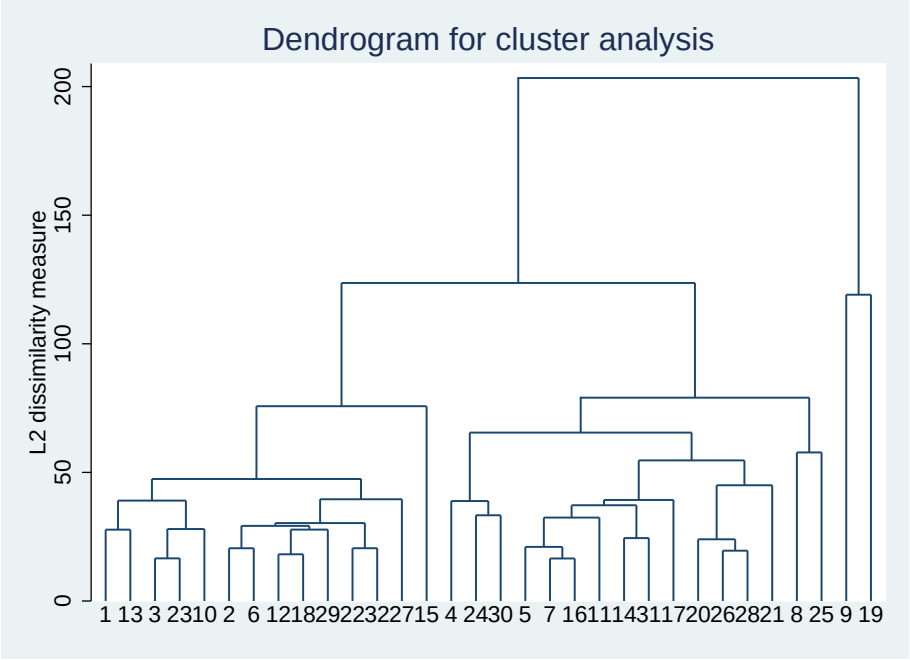
Se llevó a cabo de nuevo el análisis eliminando al Estado de México, el cual se une al fnal, y se obtuvieron los siguientes resultados:



Se presenta de nuevo el efecto de *encadenamiento*, lo cual complica la conformación de las regiones. Nótese que los grupos de estados se conservan, incluso se observa que el estado de Hidalgo (13) ya se separa ligeramente de ambos grupos, aunque por su cercanía, se le puede considerar como parte de la región 1. Se forman entonces las siguientes regiones:

Región 1	Aguascalientes, Oaxaca, Tamaulipas, Coahuila, Morelos, Nayarit, Chiapas, Michoacán, Tlaxcala, Sonora, Jalisco, Yucatán, Guerrero, Querétaro, Zacatecas, Baja California, Guanajuato, Puebla, (Hidalgo).
Región 2	Baja California Sur, Quintana Roo, Colima, Chihuahua, San Luis Potosí, Distrito Federal, Durango, Veracruz, Nuevo León, Campeche, Sinaloa, Tabasco.

4.3.1.5. Análisis de Conglomerados con *liga peso - promedio ponderado* (Media de grupos), tomando como medida de disimilaridad la norma euclideana.

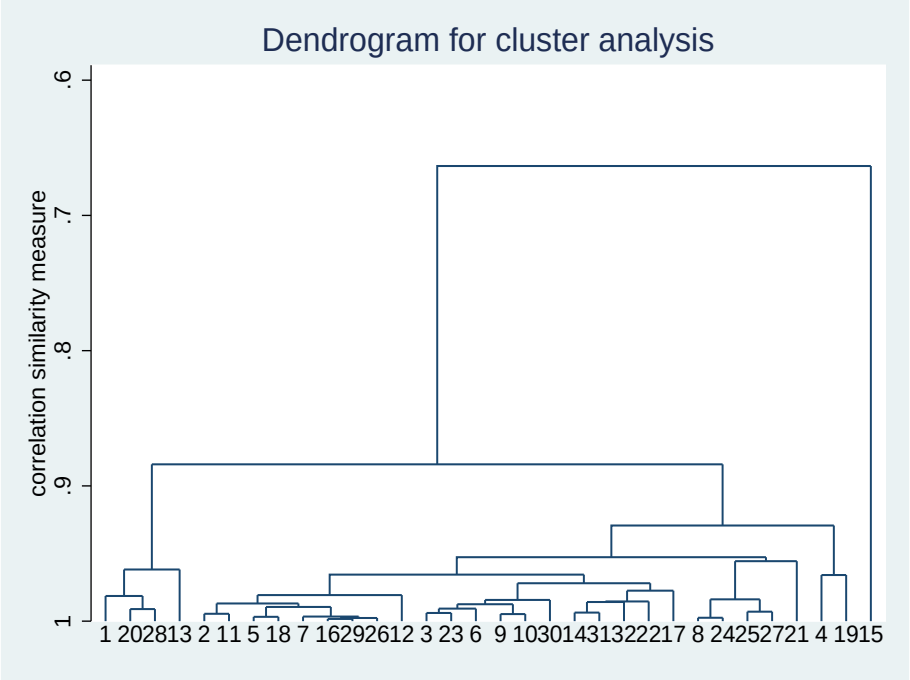


Realizando el corte correspondiente, se obtienen las siguientes regiones:

Región 1	(Aguascalientes, Hidalgo, Baja California, Colima, Guerrero, Nayarit, Tlaxcala, Querétaro, Zacatecas, Tabasco, Quintana Roo, Durango, Baja California Sur.
Región 2	México).
Región 3	Campeche, San Luis Potosí, Veracruz, Coahuila, Chiapas, Michocán, Jalisco, Yucatán, Guanajuato, Morelos, Sonora, Tamaulipas, Puebla, Oaxaca.
Región 4	Chihuahua, Sinaloa.
Región 5	Distrito Federal, Nuevo León.

Se observan regiones concretas al usar como medida de disimilaridad la norma euclideana. El estado de México ha mantenido su comportamiento, aunque no tan claro como casos previos, por ello bien formaría parte de la región uno, como se muestra con los paréntesis.

4.3.1.6. Análisis de Conglomerados con *liga peso - promedio ponderado* (Media de grupos), tomando como medida de similaridad la matriz de correlaciones.

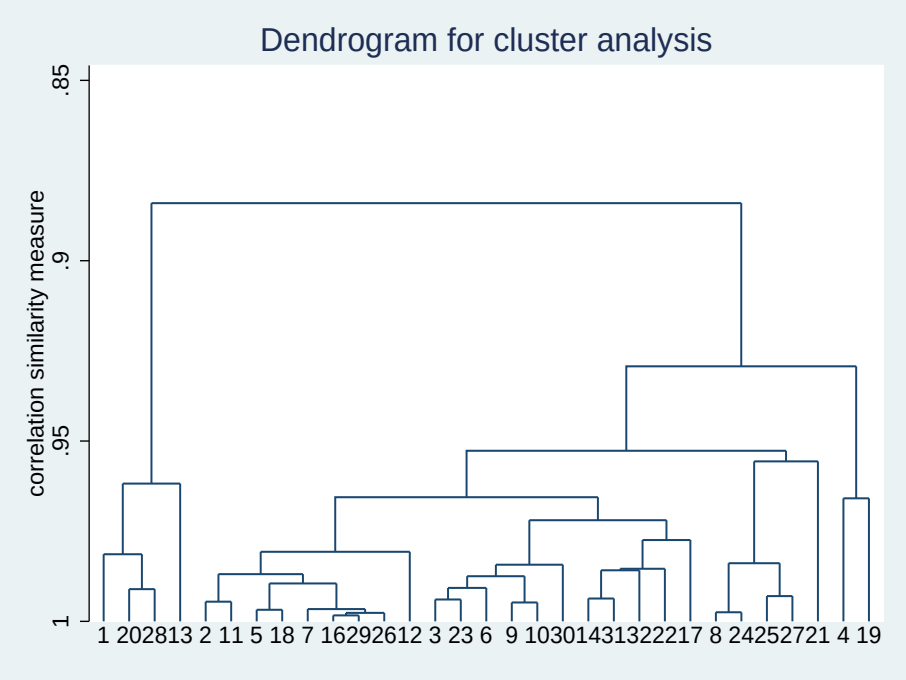


Al realizar el corte correspondiente respecto a la medida de similaridad, se obtienen las siguientes regiones:

Región 1	Aguascalientes, Oaxaca, Tamaulipas, Hidalgo.
Región 2	(Baja California, Guanajuato, Coahuila, Nayarit, Chiapas, Michoacán, Tlaxcala, Sonora, Guerrero), (Baja California Sur, Quintana Roo, Colima, Distrito Federal, Durango, Veracruz, Jalisco, Yucatán, Puebla, Querétaro, Chihuahua), Sinaloa, San Luis Potosí, Zacatecas, Tabasco, Morelos.
Región 4	Campeche, Nuevo León.
Región 5	México.

Nótese el efecto de *encadenamiento* en el dendrograma, lo cual no permite determinar regiones más claras hacia el centro del mismo.

Nuevamente, el estado de México se une al final en el proceso de formación del dendrograma. Al eliminarlo y rehacer el anásis se obtiene:

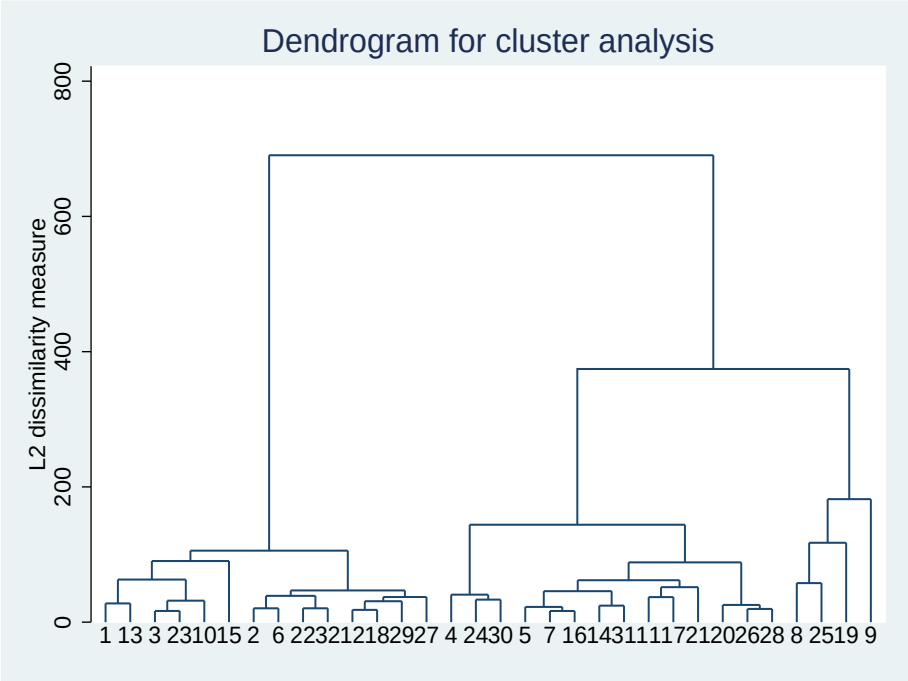


Se conserva el comportamiento de Campeche (4) y Nuevo León (19), así como la primera región (Aguascalientes, Oaxaca, Tamaulipas e Hidalgo). Sin embargo, es al centro donde posiblemente haya una partición más fina[14] que la anterior, pero más adelante se observará en la gráfica de componentes principales (4.3.4) que se puede considerar como un solo grupo. De manera que se tiene:

Región 1	Aguascalientes, Oaxaca, Tamaulipas, Hidalgo.
Región 2	(Baja California, Guanajuato, Coahuila, Nayarit, Chiapas, Michoacán, Tlaxcala, Sonora, Guerrero), (Baja California Sur, Quintana Roo, Colima, Distrito Federal, Durango, Veracruz, Jalisco, Yucatán, Zacatecas, Querétaro, Morelos), (Chihuahua, San Luis Potosí, Sinaloa, Tabasco, Puebla).
Región 4	Campeche, Nuevo León.

[14] Partición en tres regiones para un total de cinco

4.3.1.7. Análisis de Conglomerados con el *Método de Ward*, tomando como medida de disimilaridad la norma euclideana.

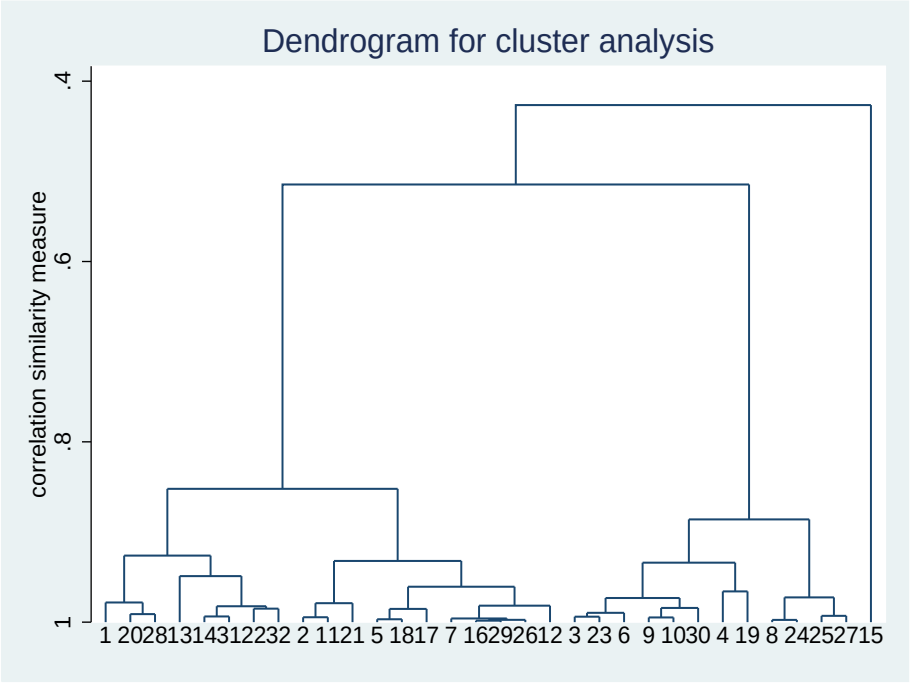


Realizando el corte correspondiente respecto a la medida de disimilaridad se obtienen las siguientes regiones:

Región 1	Aguascalientes, Hidalgo, Baja California Sur, Quintana Roo, Durango, México, Baja California, Colima, Querétaro, Zacatecas, Guerrero, Nayarit, Tlaxcala, Tabasco.
Región 2	Campeche, San Luis Potosí, Veracruz, Coahuila, Chiapas, Michoacán, Jalisco, Yucatán, Guanajuato,Morelos, Puebla, Oaxaca, Sonora, Tamaulipas.
Región 3	Chiapas, Sinaloa, Nuevo León, Distrito Federal.

Las regiones 1 y 2 se han conservado en los análisis 4.3.1.3 y 4.1.3.5. Se presenta levemente el efecto de *encadenamiento* y por ello no es tan clara la "*separación*" del Distrito Federal.

4.3.1.8. Análisis de Conglomerados con el *Método de Ward*, tomando como medida de similaridad la matriz de correlaciones.

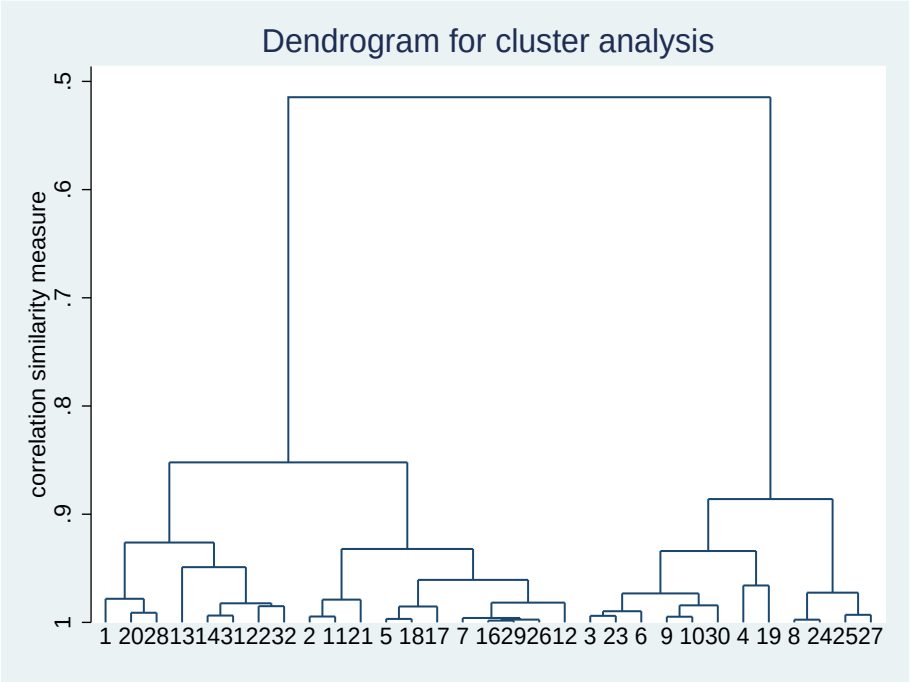


Al realizar el corte correspondiente respecto a la medida de similaridad se obtienen las siguientes regiones:

Región 1	Aguascalientes, Oaxaca, Tamaulipas, Hidalgo, Jalisco, Yucatán, Querétaro, Zacatecas.
Región 2	Baja California, Guanajuato, Puebla, Coahuila, Nayarit, Morelos, Chiapas, Michoacán, Tlaxcala, Sonora, Guerrero.
Región 3	Baja California Sur, Quintana Roo, Colima, Distrito Federal, Durango, Veracruz, Campeche, Nuevo León, Chihuahua, San Luis Potosí, Sinaloa, Tabasco.
Región 4	México.

Con base en el dendrograma, pudieron haber sido sólo tres regiones, donde una de éstas estaría formada únicamente por el estado de México.

Se eliminó al estado de México para realizar nuevamente el anásis resultando las mismas regiones y solamente se presentan cambios de forma en el dendrograma:



Se conservan los primeros cuatro grupos, cuyo comportamiento ya se percibe de manera más clara. Aunque se puede considerar que únicamente se tiene un par de conjuntos [15]:

Región 1	Aguascalientes, Oaxaca, Tamaulipas, Hidalgo, Jalisco, Yucatán, Querétaro, Zacatecas.
Región 2	Baja California, Guanajuato, Puebla, Coahuila, Nayarit, Morelos, Chiapas, Michoacán, Tlaxcala, Sonora, Guerrero.
Región 3	Baja California Sur, Quintana Roo, Colima, Distrito Federal, Durango, Veracruz, Campeche, Nuevo León.
Región 4	Chihuahua, San Luis Potosí, Sinaloa, Tabasco.

[15] donde el primero se conforma por la unión de las regiones uno y dos, y el segundo por la unión de las regiones tres y cuatro.

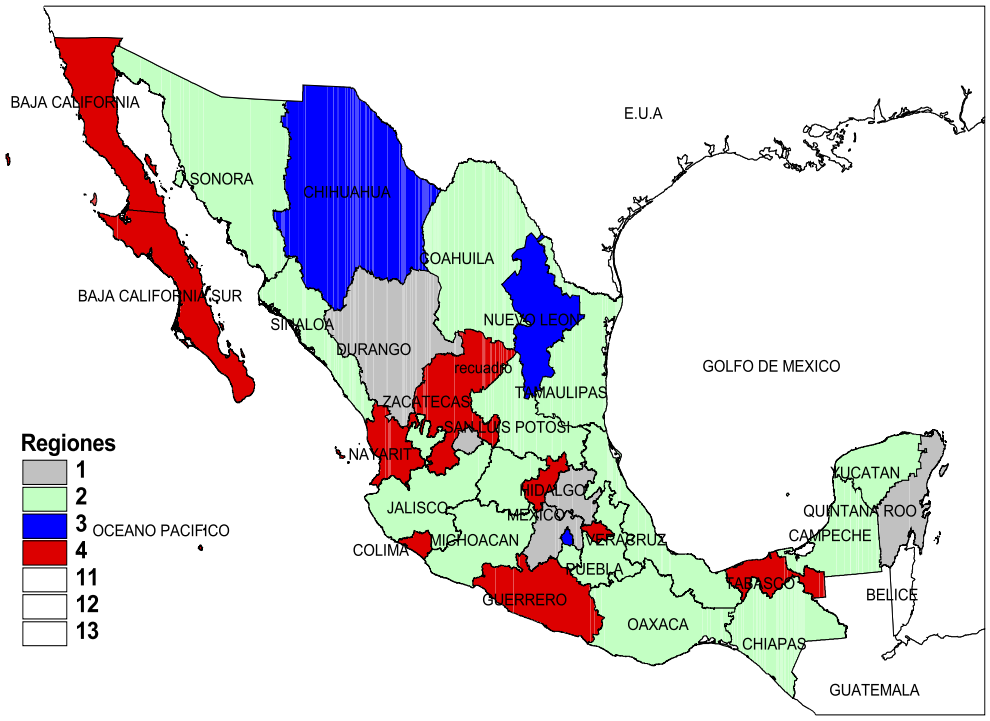
4.3.2. Regiones obtenidas con el Algoritmo de Partición K - medias.

Las siguientes, son las regiones obtenidas con el algoritmo de partición clásico de las k-medias para 4 y 5 regiones.

- **Tomando como medida de disimilaridad la norma euclideana y k observaciones aleatorias como centros iniciales de grupos.**

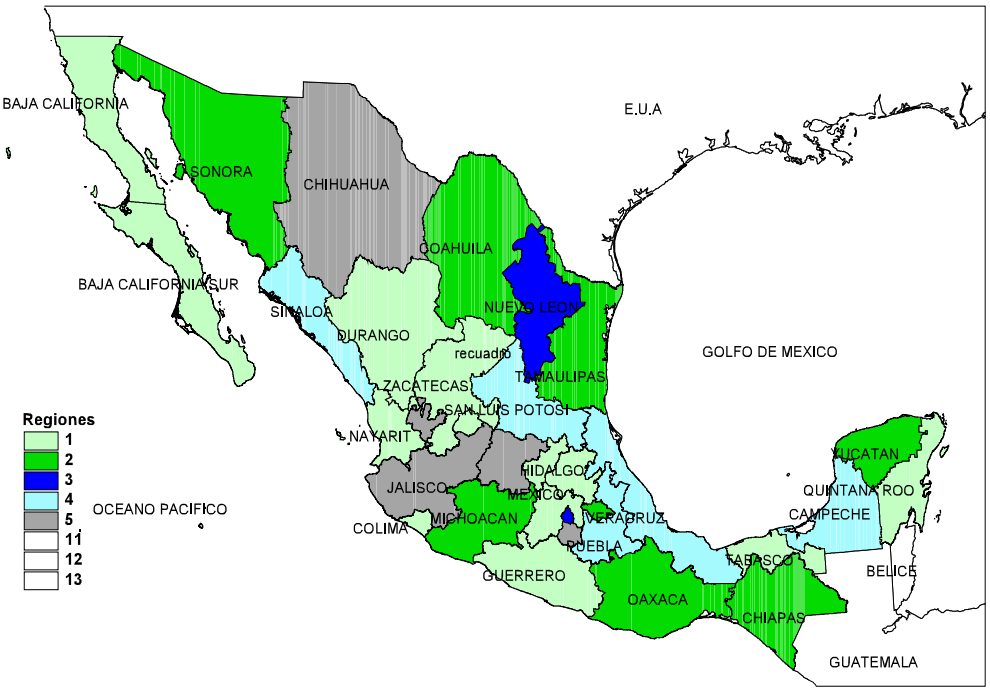
Caso I. k=4 regiones.

Región 1	Aguascalientes, Durango, Hidalgo, México, Quintana Roo.
Región 2	Campeche, Coahuila, Chiapas, Guanajuato, Jalisco, Michoacán, Morelos, Oaxaca, Puebla, San Luis Potosí, Sinaloa, Sonora, Tamaulipas, Veracruz, Yucatán.
Región 3	Chihuahua, Distrito Federal, Nuevo León.
Región 4	Baja California, Baja California Sur, Colima, Guerrero, Nayarit, Querétaro, Tabasco, Tlaxcala, Zacatecas.



Caso II. k=5 regiones.

Región 1	Aguascalientes, Baja California, Baja California Sur, Colima, Durango, Guerrero, Hidalgo, México, Nayarit, Querétaro, Tabasco, Quintana Roo, Zacatecas.
Región 2	Coahuila, Chiapas, Michoacán, Oaxaca, Sonora, Tamaulipas, Tlaxcala, Yucatán.
Región 3	Distrito Federal, Nuevo León.
Región 4	Campeche, Puebla, San Luis Potosí, Sinaloa, Veracruz.
Región 5	Chihuahua, Guanajuato, Jalisco, Morelos.



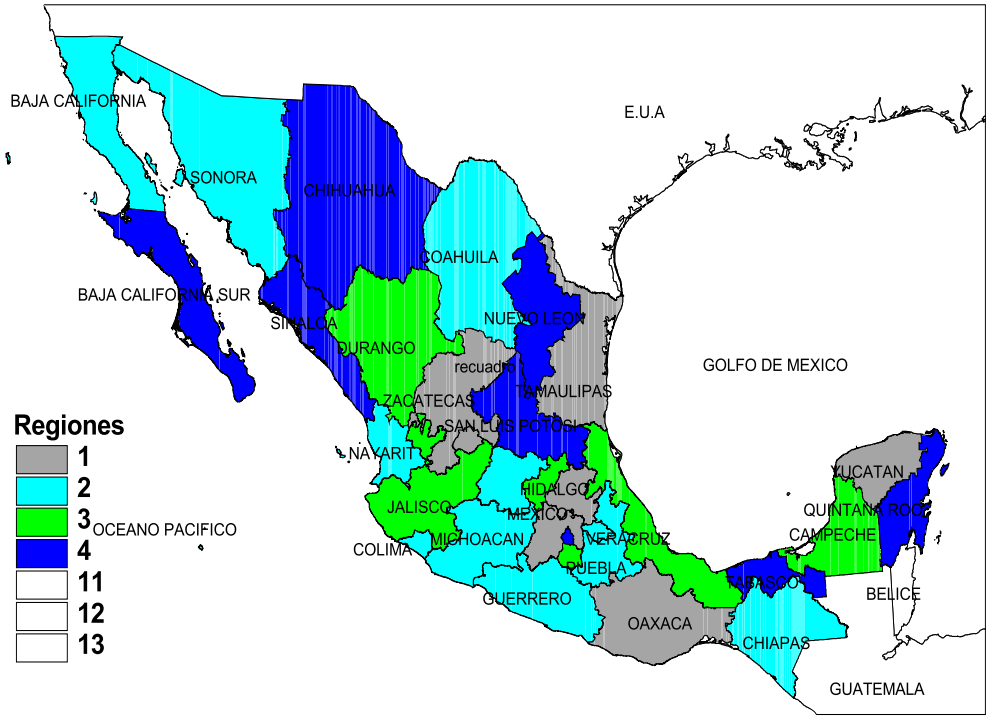
Observaciones:

La unión de las regiones 1 y 4 del **Caso I**, proporciona como resultado la **Región 1** del **Caso II**. Las regiones 4 y 5 de **Caso II**, conforman en esencia la **Región 1** del primer caso. Paralelamente se puede notar que el estado de Nuevo León y el Distrito Federal tienen un comportamiento similar en los dos análisis.

- **Tomando como medida de similaridad la matriz de correlaciones y k observaciones aleatorias como centros iniciales de grupos.**

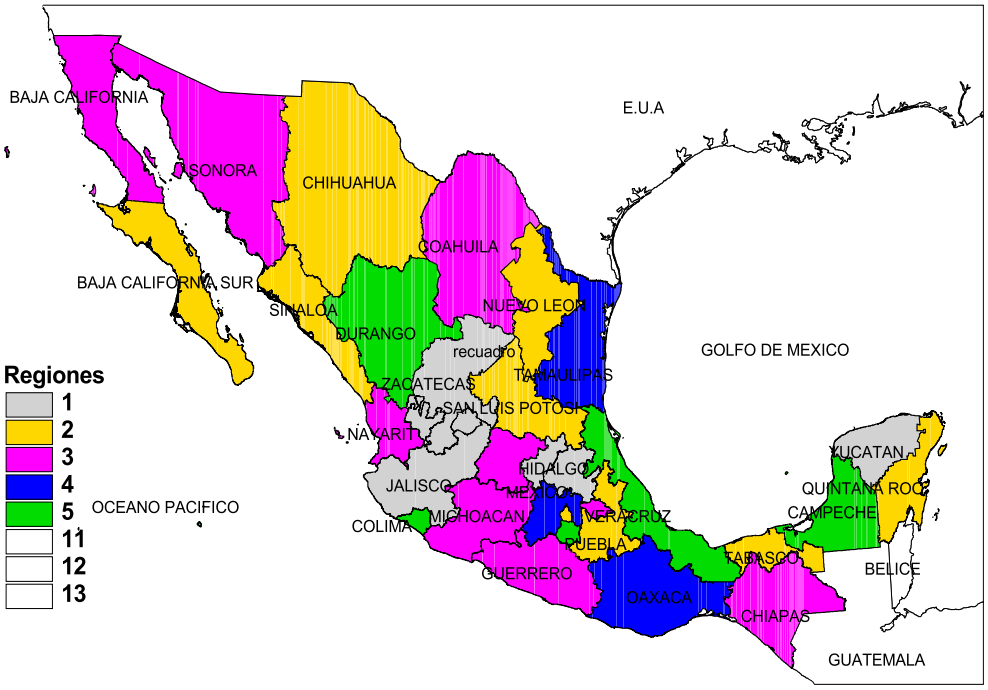
Caso I. k=4 regiones.

Región 1	Aguascalientes, Hidalgo, México, Oaxaca, Tamaulipas, Yucatán, Zacatecas.
Región 2	Baja California, Coahuila, Colima, Chiapas, Guanajuato, Guerrero, Michoacán, Nayarit, Puebla, Sonora, Tlaxcala.
Región 3	Campeche, Durango, Jalisco, Morelos, Querétaro, Veracruz.
Región 4	Baja California Sur, Chihuahua, Distrito Federal, Nuevo León, Quintana Roo, San Luis Potosí, Sinaloa, Tabasco.



Caso II. k=5 regiones.

Región 1	Aguascalientes, Hidalgo, Jalisco, Querétaro, Yucatán, Zacatecas.
Región 2	Baja California Sur, Chihuahua, Distrito Federal, Nuevo León, Puebla, Quintana Roo, San Luis Potosí, Sinaloa, Tabasco.
Región 3	Baja California, Coahuila, Chiapas, Guanajuato, Guerrero, Nayarit, Michoacán, Sonora, Tlaxcala.
Región 4	México, Oaxaca, Tamaulipas.
Región 5	Campeche, Colima, Durango, Morelos, Veracruz.

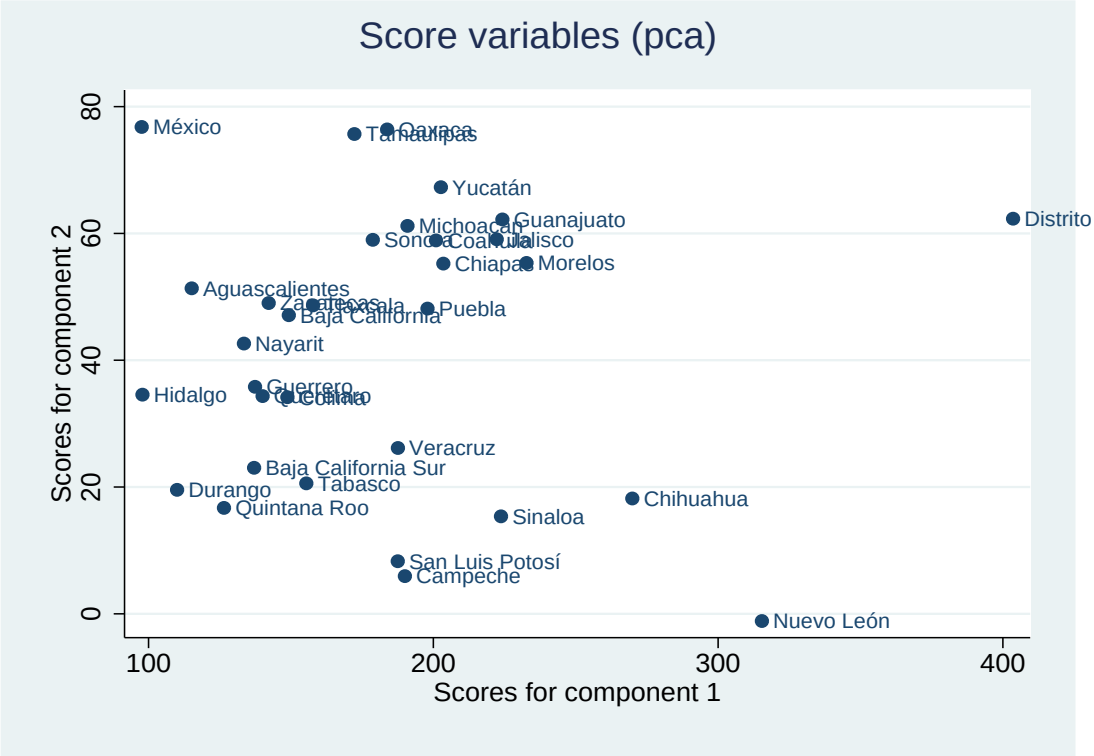


Observaciones:

1. La unión de las regiones tres y cuatro del **Caso II**, conforma la **Región 2** del **Caso I**.

2. Los estados de Durango, Jalisco, Nuevo León y Tlaxcala se conservan en una misma región en los dos casos.

4.3.3. Gráfica de los dos primeros componentes principales obtenidos con la matriz de varianzas y covarianzas.



El porcentaje de varianza explicado por las primeras dos componentes es 90 %.

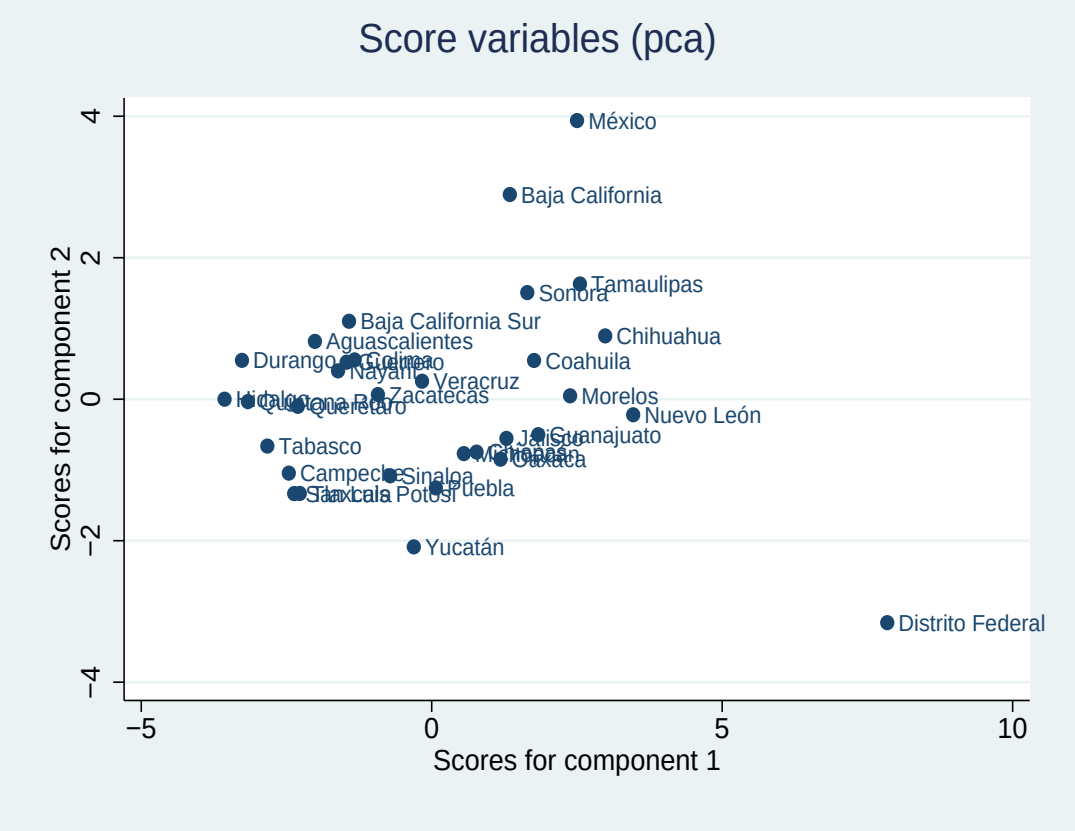
Observaciones:

1. De inmediato se nota el comportamiento del Distrito Federal, que no se había detectado previamente.

2. Asimismo, se tienen los "*alejamientos*" de los estados de México y Nuevo León.

3. A la derecha se tiene un grupo conformado principalmente por Hidalgo, Nayarit, Guerrero y Baja California, entre otros, que se mantuvo durante los análisis.

4.3.4. Gráfica de los dos primeros componentes principales obtenidos con la matriz de correlaciones.



El porcentaje de varianza explicado por las primeras dos componentes es 68 %.

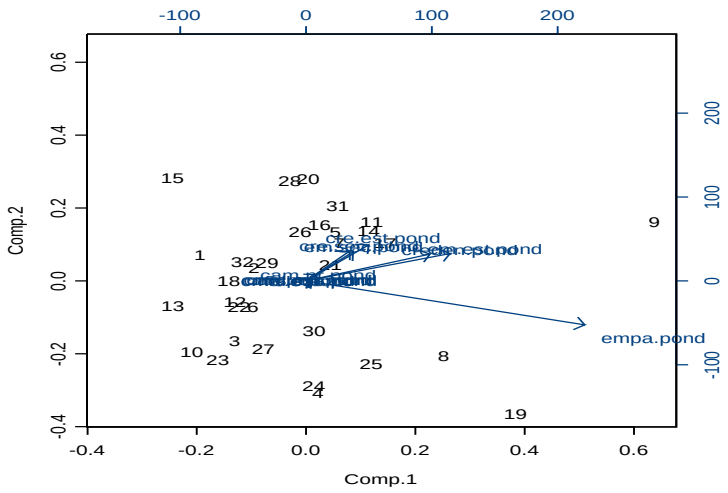
Observaciones:

1. Nuevamente se observa un comportamiento del Distrito Federal que no se había detectado con las análisis previos. Además se corrobora el "*alejamiento*" de los estados de Baja California y México.

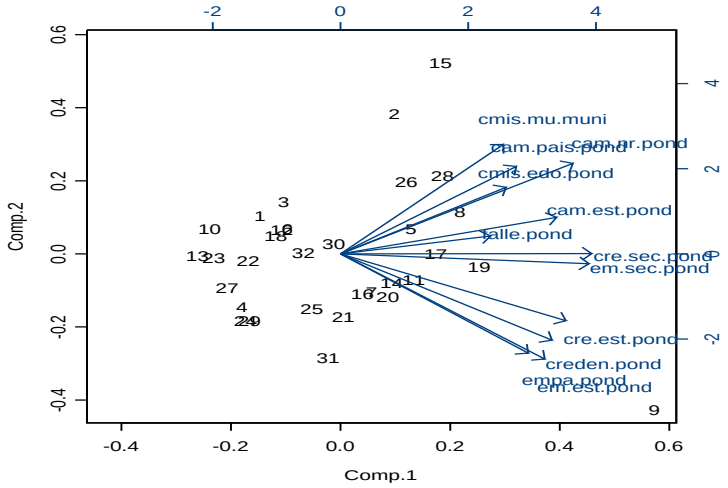
2. Al centro se tiene un grupo importante de donde se obtiene principalmente, dos grupos densos que se habían detectado.

4.3.5. Gráficas Biplot de los componentes principales obtenidos con los indicadores ponderados.

4.3.5.1. Biplot con la matriz de varianzas y covarianzas.



4.3.5.2. Biplot con la matriz de correlaciones.



En la primera gráfica se observa la poca relación del Distrito Federal (9) y el estado de México (15) con los restantes; el indicador más influyente es Empadronados. En la segunda gráfica nótese nuevamente al estado de México (15) y

a Baja California (2) quienes ahora son los menos relacionados con las restantes entidades y presentan un *alejamiento* que ya había sido detectado.

En este análisis (VNM2005 con indicadores ponderados), el comportamiento de las entidades federativas fue errático, a diferencia de las secciones previas 4.1 y 4.2, en donde se analizaron los indicadores sin ponderar. En efecto, ahora se pudo notar una clara diferencia en cuanto a resultados. Además, en la mayoría de los éstos se reportan cinco regiones, pero donde dos o tres de éstas se conforman a lo más, por tres entidades (por ejemplo subsecciones 4.3.1.3, 4.3.1.5, 4.3.1.6.). Solamente el comportamiento *aislado* del estado de México puede apoyarse. Asimismo, en todos los casos es posible demarcar dos macro - regiones. En este sentido, no es posible exhibir regiones *consistentes* en cuanto a entidades federativas que las conforman. Veamos ahora lo que ocurre con la VNM2006.

4.3.6. Resultados con los indicadores ponderados para VNM2006

Ahora se muestran los resultados obtenidos con la VNM2006. Los indicadores se ponderaron con el recíproco de su desviación estándar estimada. Se exhiben:

1. Los dendrogramas resultantes del Análisis de Conglomerados y las regiones formadas al proceder con los cortes correpondientes.

2. Las regiones obtenidas con el Algoritmo de Partición k-medias para cuatro y cinco clusters[16].

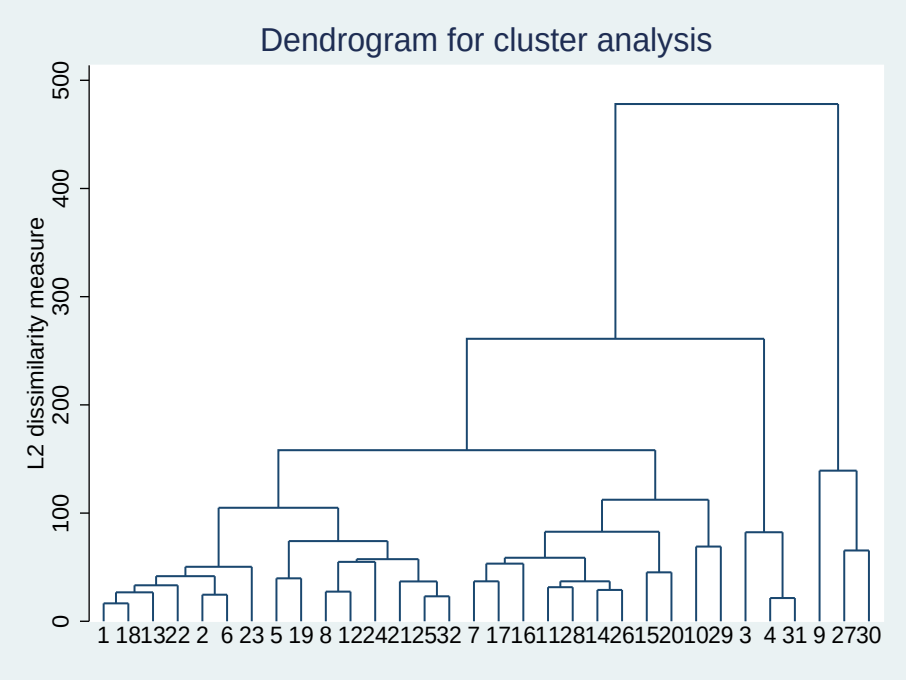
3. Gráfica de los dos primeros componentes principales obtenidos con la matriz de varianzas y covarianzas.

4. Gráfica de los dos primeros componentes principales obtenidos con la matriz de correlaciones.

5. Biplots suponiendo homoscedasticidad.

[16] Siguiendo la idea de conservar la estructura obtenida.

4.3.6.1. Análisis de Conglomerados con *liga completa*, tomando como medida de disimilaridad la norma euclideana.

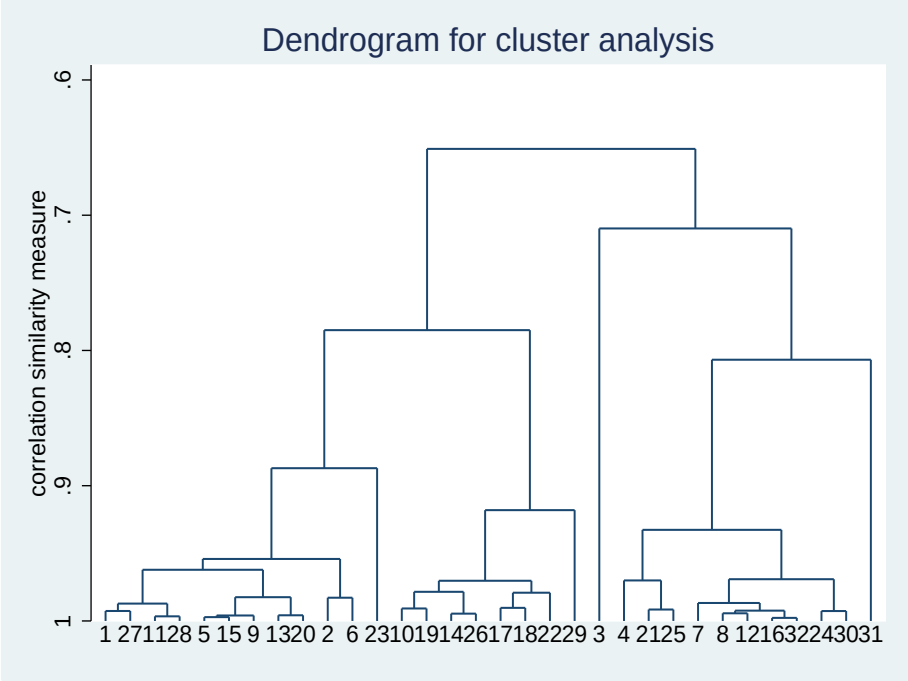


Realizando el corte correspondiente se obtienen las siguientes regiones:

Región 1	Aguascalientes, Nayarit, Hidalgo, Querétaro, Baja California, Colima, Quintana Roo, Coahuila, Nuevo León, Chihuahua, San Luis Potosí, Guerrero, Puebla, Sinaloa, Zacatecas.
Región 2	Chiapas, Morelos, Michoacán, Guanajuato, Tamaulipas, Jalisco, Sonora, México, Oaxaca, Durango, Tlaxcala.
Región 3	Baja California Sur, Campeche, Yucatán.
Región 4	Distrito Federal, Tabasco, Veracruz.

Ahora los estados de Campeche y Yucatán se "*unen*" a Baja California Sur (que había estado apartado del resto) y se observa la "separación" del Distrito Federal junto con Tabasco y Veracruz.

4.3.6.2. Análisis de Conglomerados con *liga completa*, tomando como medida de similaridad la matriz de correlaciones.

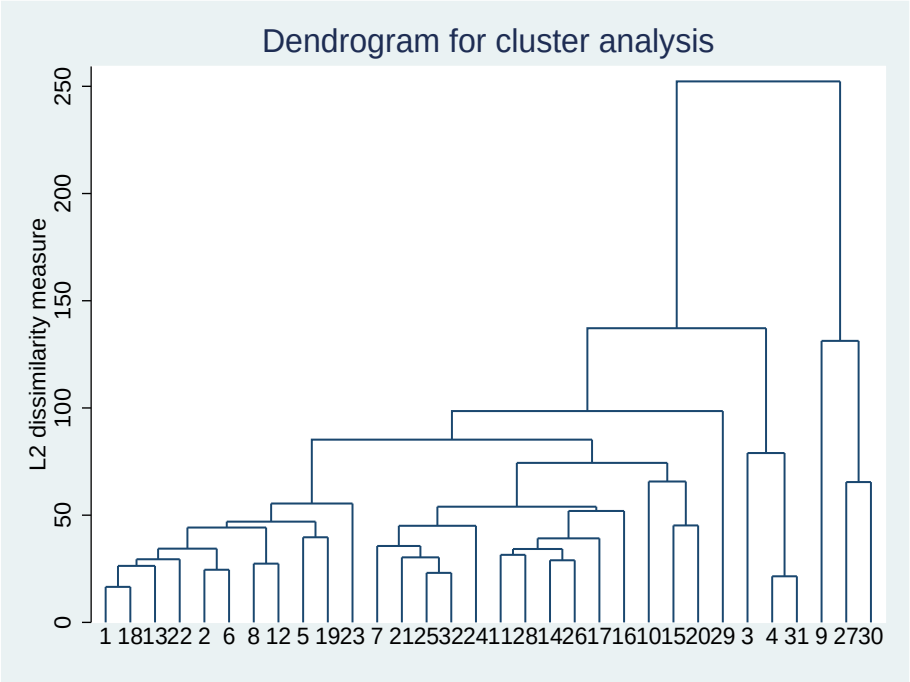


Al realizar el corte correspondiente respecto a la medida de similaridad, se obtienen las siguientes regiones:

Región 1	Aguascalientes, Tabasco, Guanajuato, Tamaulipas, Coahuila, México, Distrito Federal, Hidalgo, Oaxaca, Baja California, Colima, Quintana Roo.
Región 2	Durango, Nuevo León, Jalisco, Sonora, Morelos, Querétaro, Nayarit, Tlaxcala .
Región 3	Baja California Sur.
Región 4	(Campeche, Puebla, Sinaloa, Chiapas, Chihuahua, Guerrero, Michoacán, Zacatecas, San Luis Potosí, Veracruz.
Región 5	Yucatán.)

Nótese que el estado de Quintana Roo ya se encuentra en la **Región** 1, y separado de Baja California Sur. De hecho, pueden considerarse cuatro regiones, uniendo las dos últimas, como se muestra en el cuadro.

4.3.6.3. Análisis de Conglomerados con *liga promedio*, tomando como medida de disimilaridad la norma euclideana.

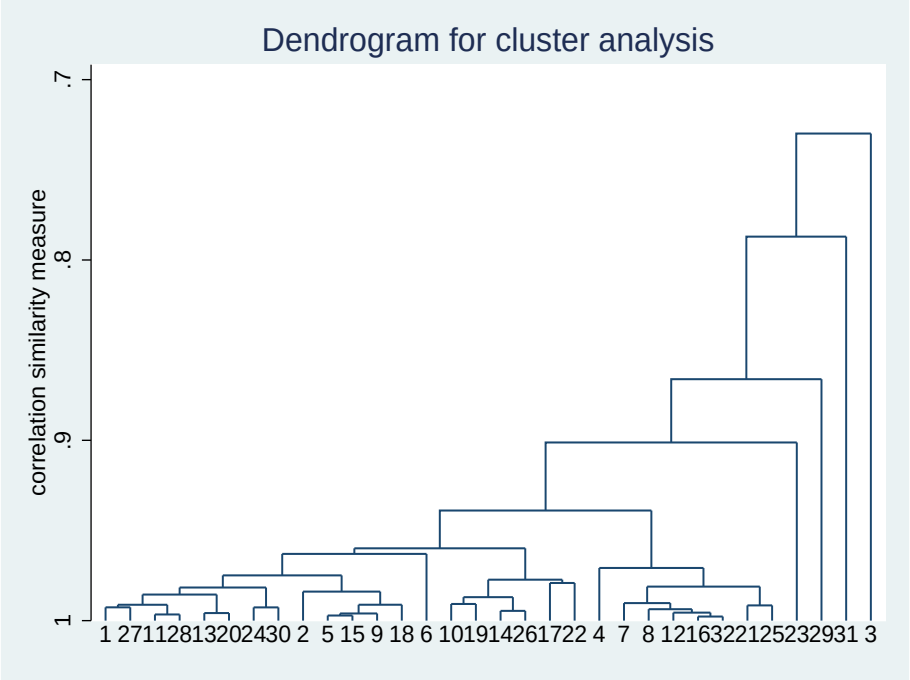


Al ejecutar el corte correspondiente respecto a la medida de disimilaridad, resultan las siguientes regiones:

Región 1	Aguascalientes, Nayarit, Hidalgo, Querétaro, Baja California, Colima, Quintana Roo, Coahuila, Nuevo León, Chihuahua, Guerrero.
Región 2	Chiapas, Morelos, Michoacán, Guanajuato, Tamaulipas, Jalisco, México, Oaxaca, Durango, Tlaxcala, Puebla, Sinaloa, Zacatecas, Sonora, San Luis Potosí.
Región 3	Baja California Sur, Campeche, Yucatán.
Región 4	Distrito Federal, Tabasco, Veracruz.

Respecto al análisis realizado con la *liga completa* se conservan las regiones 3 y 4. Y aunque con algunos cambios, también las regiones 1 y 2 se mantienen.

4.3.6.4. Análisis de Conglomerados con *liga promedio*, tomando como medida de similaridad la matriz de correlaciones.

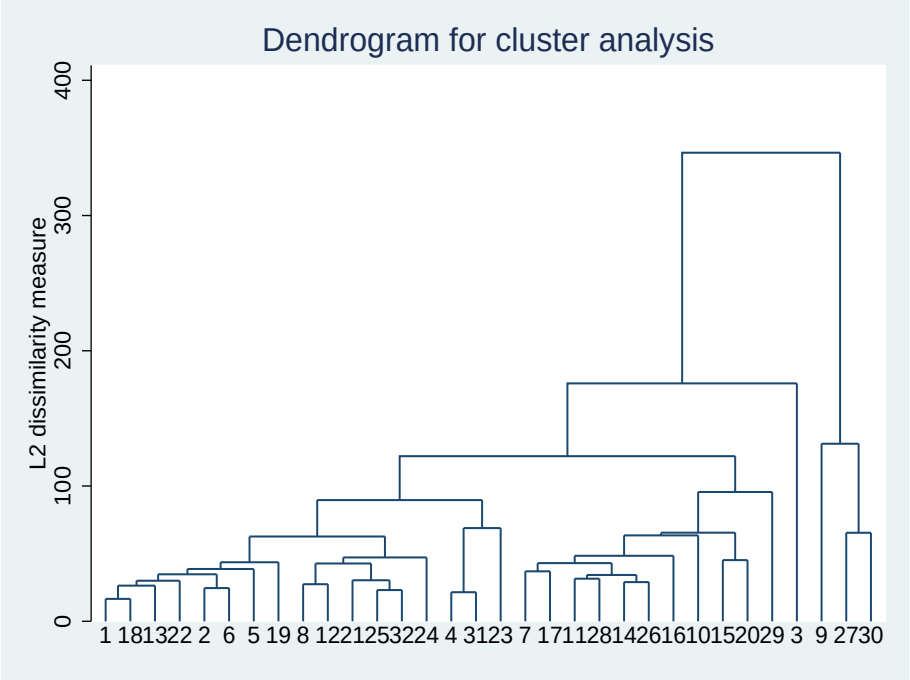


Con el corte correspondiente, se obtienen las siguientes regiones:

Región 1	Aguascalientes, Tabasco, Guanajuato, Tamaulipas, Hidalgo, Oaxaca, San Luis Potosí, Veracruz, Baja California, Coahuila, México, Distrito Federal, Nayarit, Colima, Durango, Nuevo León, Jalisco, Sonora, Morelos, Querétaro.
Región 2	Campeche, Chiapas, Chihuahua, Guerrero, Michoacán, Zacatecas, Puebla, Sinaloa.
Región 3	Quintana Roo.
Región 4	Tlaxcala.
Región 5	Yucatán.
Región 6	Baja California Sur.

Se presenta un efecto de *encadenamiento* lo cual influye en la formación del dendrograma y por ende en las regiones. La región 1 conserva su estructura.

4.3.6.5. Análisis de Conglomerados con *liga peso - promedio ponderado* (Media de grupos), tomando como medida de disimilaridad la norma euclideana.

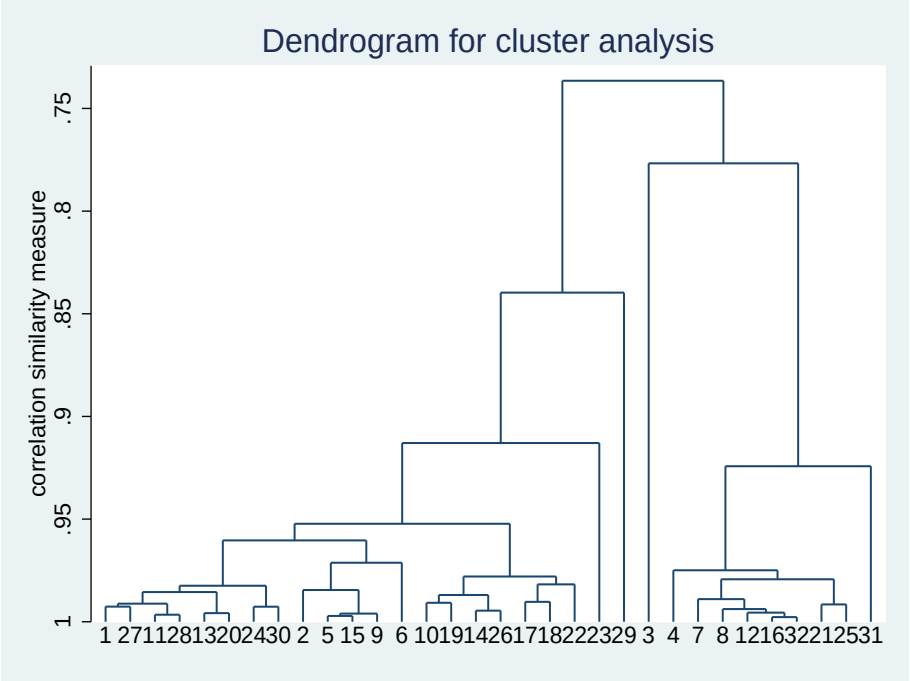


Al realizar el corte correspondiente respecto a la medida de disimilaridad, se obtienen las siguientes regiones:

Región 1	Aguascalientes, Nayarit, Hidalgo, Querétaro, Baja California, Colima, Coahuila, Nuevo León, Chihuahua, San Luis Potosí, Guerrero, Puebla, Sinaloa, Zacatecas.
Región 2	Quintana Roo, Campeche, Yucatán.
Región 3	Chiapas, Morelos, Michoacán, Guanajuato, Tamaulipas, Jalisco, Sonora, México, Oaxaca, Durango, Tlaxcala.
Región 4	Baja California Sur.
Región 5	Distrito Federal, Tabasco, Veracruz.

Se conserva la **Región 5**. Ahora, Campeche y Yucatán se "unen" a Quintana Roo, mientras que las regiones 1 y 3 se mantienen en esencia.

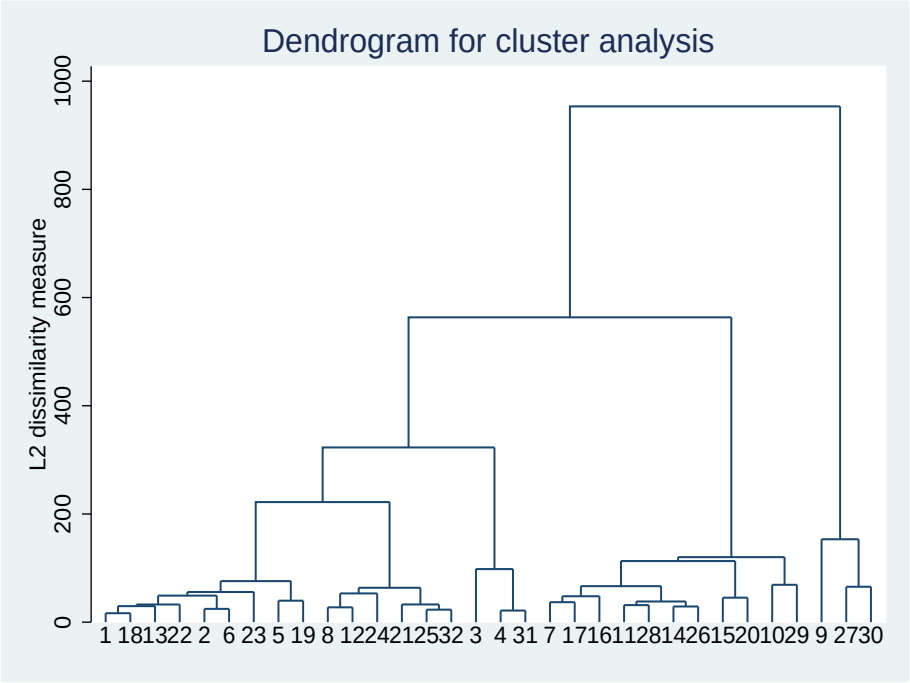
4.3.6.6. Análisis de Conglomerados con *liga peso - promedio ponderado* (Media de grupos), tomando como medida de similaridad la matriz de correlaciones.



Al ejecutar el corte correspondiente respecto a la medida de similaridad, se obtienen las siguientes regiones:

Región 1	Aguascalientes, Tabasco, Guanajuato, Tamaulipas, Hidalgo, Oaxaca, San Luis Potosí, Veracruz, Baja California, Coahuila, México, Colima Distrito Federal.
Región 2	Durango, Nuevo León, Jalisco, Sonora, Morelos, Querétaro, Nayarit.
Región 3	Campeche, Chiapas, Chihuahua, Guerrero, Michoacán, Zacatecas, Puebla, Sinaloa.
Región 4	Quintana Roo.
Región 5	Tlaxcala.
Región 6	Yucatán.
Región 7	Baja California Sur.

4.3.6.7. Análisis de Conglomerados con el *Método de Ward*, tomando como medida de disimilaridad la norma euclideana.

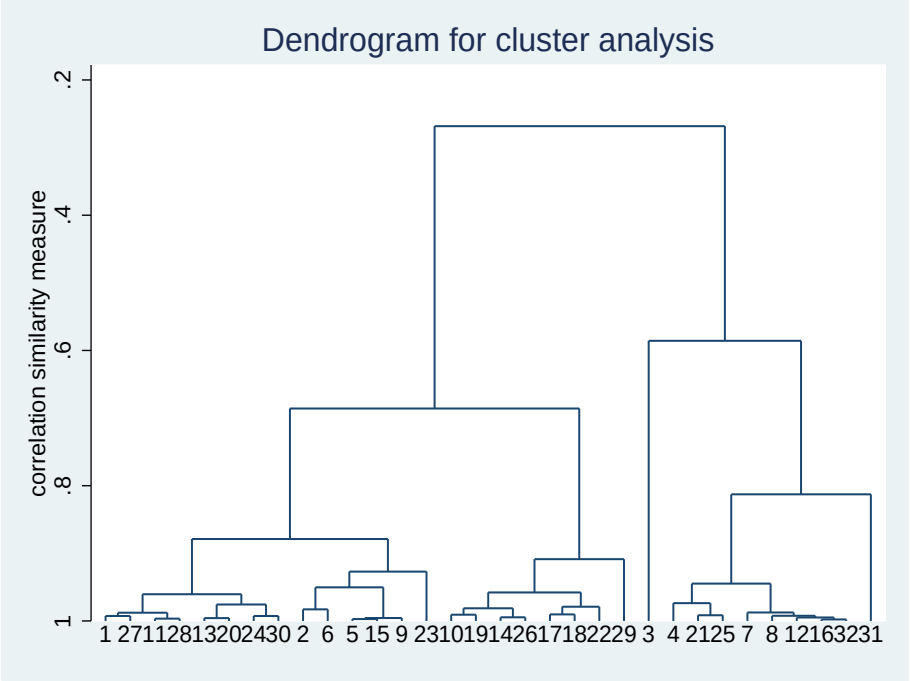


Realizando el corte correspondiente respecto a la medida de disimilaridad se obtienen las siguientes regiones:

Región 1	Aguascalientes, Nayarit, Hidalgo, Querétaro, Baja California, Colima, Quintana Roo, Coahuila, Nuevo León, Chihuahua, San Luis Potosí, Guerrero, Puebla, Sinaloa, Zacatecas.
Región 2	Chiapas, Morelos, Michoacán, Guanajuato, Tamaulipas, Jalisco, Sonora, México, Oaxaca, Durango, Tlaxcala.
Región 3	Baja California Sur, Campeche, Yucatán.
Región 4	Distrito Federal, Tabasco, Veracruz.

Resalta el hecho de que estas regiones prácticamente coinciden con la mayoría de los análisis previos realizados. Principalmente las regiones 3 y 4.

4.3.6.8. Análisis de Conglomerados con el *Método de Ward*, tomando como medida de similaridad la matriz de correlaciones.



Al realizar el corte correspondiente respecto a la medida de similaridad se obtienen las siguientes regiones:

Región 1	Aguascalientes, Tabasco, Guanajuato, Tamaulipas, Hidalgo, Oaxaca, San Luis Potosí, Veracruz, Baja California, Coahuila, México, Colima Distrito Federal, Quintana Roo.
Región 2	Durango, Nuevo León, Jalisco, Sonora, Morelos, Querétaro, Nayarit, Tlaxcala.
Región 3	Campeche, Chiapas, Chihuahua, Guerrero, Michoacán, Zacatecas, Puebla, Sinaloa.
Región 4	Yucatán.
Región 5	Baja California Sur.

Nótese que las regiones obtenidas coinciden con las obtenidas en **4.3.5.2** y sobresale el comportamiento de Baja California Sur y Yucatán.

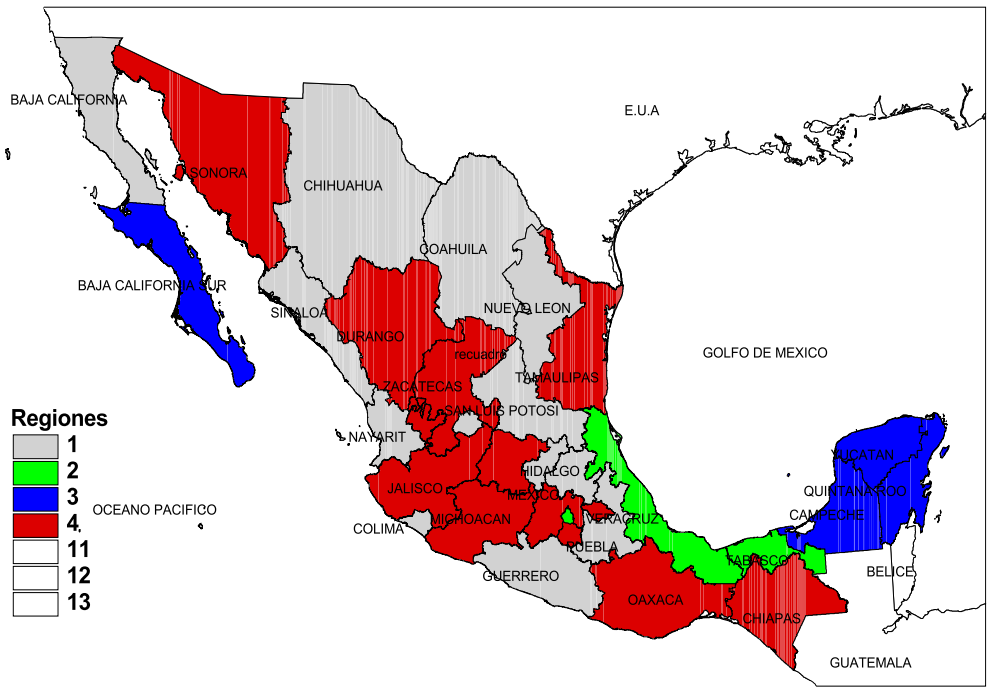
4.3.7. Regiones obtenidas con el Algoritmo de Partición K - medias.

Las siguientes, son las regiones obtenidas con el algoritmo de partición clásico de las k-medias para 4 y 5 regiones.

- **Tomando como medida de disimilaridad la norma euclideana y k observaciones aleatorias como centros iniciales de grupos.**

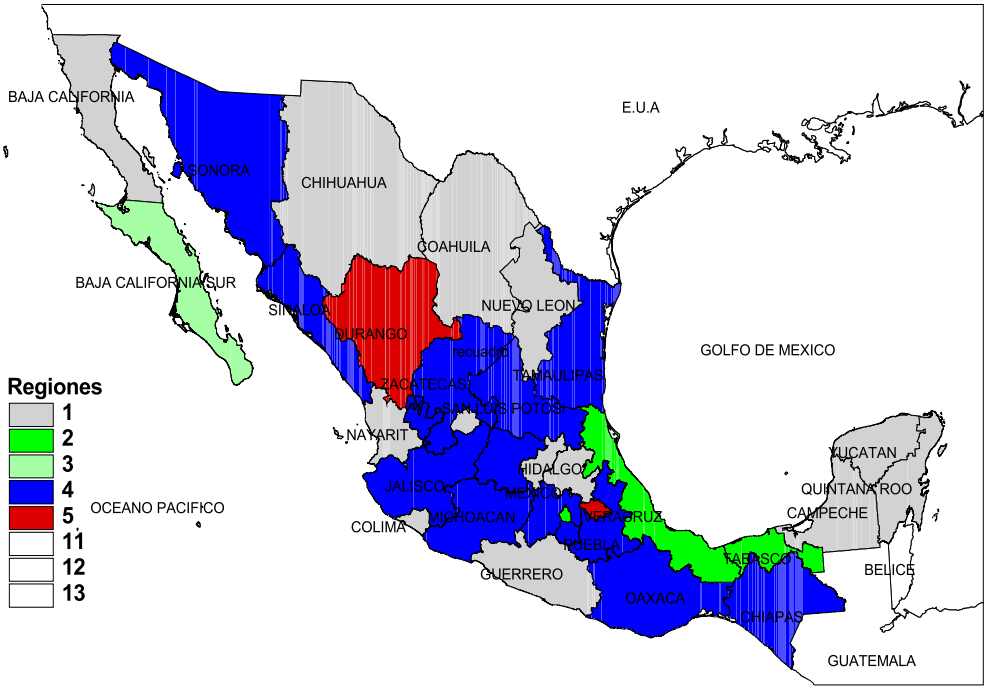
Caso I. k=4 regiones.

Región 1	Aguascalientes, Baja California, Coahuila, Colima, Chihuahua, Guerrero, Hidalgo, Nayarit, Nuevo León, Puebla, San Luis Potosí, Querétaro, Sinaloa.
Región 2	Distrito Federal, Tabasco, Veracruz.
Región 3	Baja California Sur, Campeche, Quintana Roo, Yucatán.
Región 4	Chiapas, Durango, Guanajuato, Jalisco, México, Michoacán, Morelos, Oaxaca, Sonora, Tamaulipas, Tlaxcala, Zacatecas.



Caso II. k=5 regiones.

Región 1	Aguascalientes, Baja California, Campeche, Coahuila, Colima, Chihuahua, Guerrero, Hidalgo, Nayarit, Nuevo León, Querétaro, Quintana Roo, Yucatán.
Región 2	Distrito Federal, Tabasco, Veracruz.
Región 3	Baja California Sur.
Región 4	Chiapas, Guanajuato, Jalisco, México, Michoacán, Morelos, Oaxaca, Puebla, San Luis Potosí, Sinaloa, Sonora, Tamaulipas, Zacatecas.
Región 5	Durango, Tlaxcala.



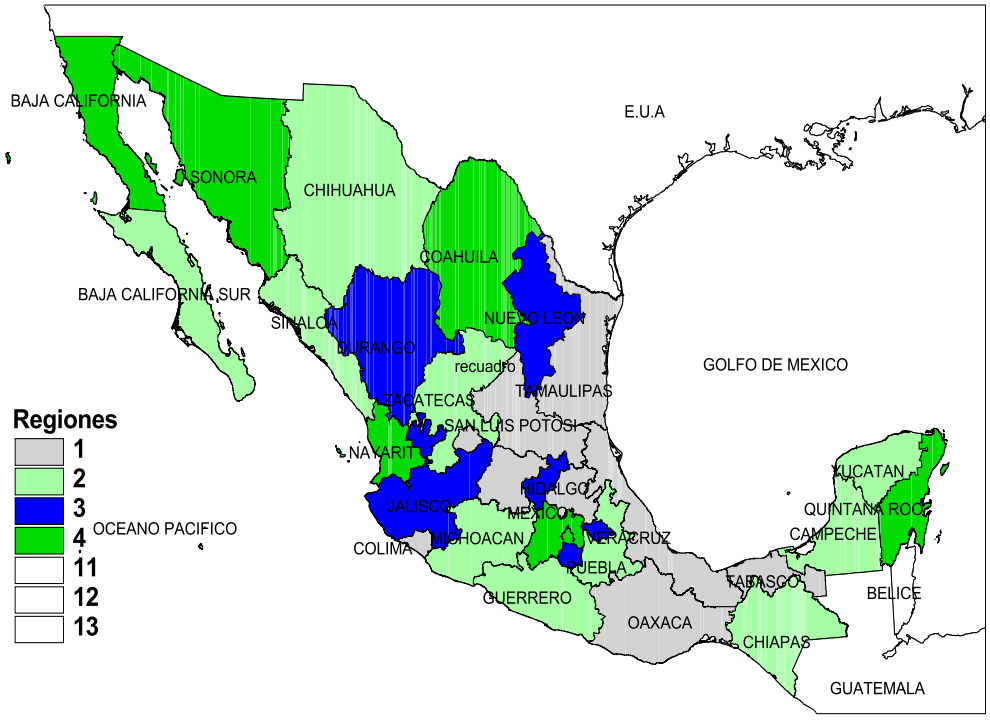
Observaciones:

Resalta la coherencia del **Caso I** con gran parte de los análisis realizados en esta sección. Por otro lado, los estados de Tabasco, Veracruz y Distrito Federal, coinciden en ambos casos en una misma región. La "novedad" es Quintana Roo, que se *une* a Yucatán y otros estados.

- **Tomando como medida de similaridad la matriz de correlaciones y k observaciones aleatorias como centros iniciales de grupos.**

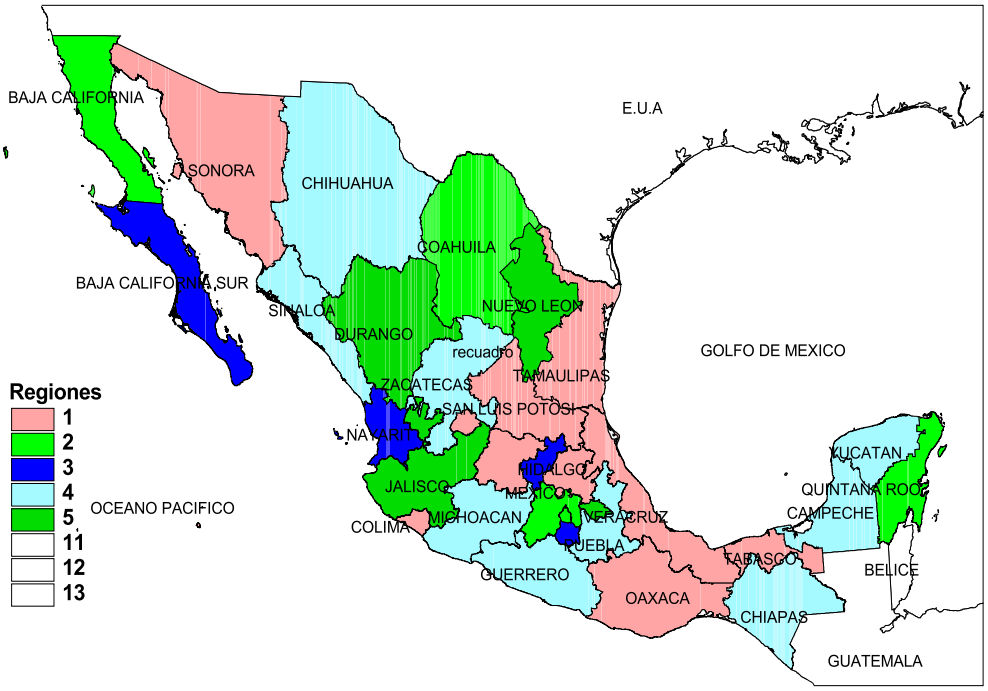
Caso I. k=4 regiones.

Región 1	Aguascalientes, Colima, Guanajuato, Hidalgo, Oaxaca, San Luis Potosí, Tabasco, Tamaulipas, Veracruz.
Región 2	Baja California Sur, Campeche, Chiapas, Chihuahua, Guerrero, Michocán, Puebla, Sinaloa, Yucatán, Zacatecas.
Región 3	Durango, Jalisco, Morelos, Nuevo León, Querétaro, Tlaxcala.
Región 4	Baja California, Coahuila, Distrito Federal, México, Nayarit, Sonora, Quintana Roo.



Caso II. k=5 regiones.

Región 1	Aguascalientes, Colima, Guanajuato, Hidalgo, Oaxaca, San Luis Potosí, Sonora, Tabasco, Tamaulipas, Veracruz.
Región 2	Baja California, Coahuila, Distrito Federal, México, Quintana Roo.
Región 3	Baja California Sur, Morelos, Nayarit, Querétaro.
Región 4	Campeche, Chiapas, Chihuahua, Guerrero, Michoacán, Puebla, Sinaloa, Yucatán, Zacatecas.
Región 5	Durango, Jalisco, Nuevo León, Tlaxcala .

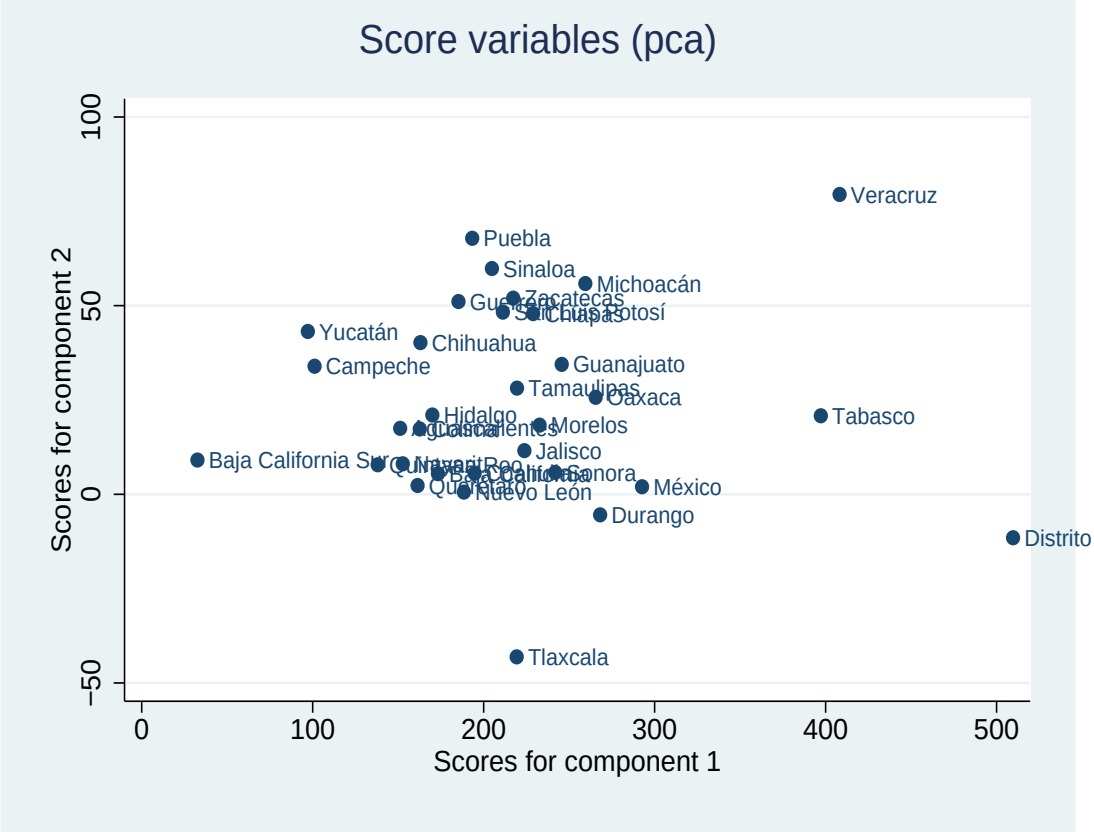


Observaciones:

1. La unión de las regiones 3 y 4 del **Caso II**, da como resultado la **Región 2** del **Caso I**.

2. Los estados de Durango, Jalisco, Nuevo León y Tlaxcala se conservan en una misma región en los dos casos.

4.3.8. Gráfica de los dos primeros componentes principales obtenidos con la matriz de varianzas y covarianzas.



El porcentaje de varianza explicado por las primeras dos componentes es 94 %.

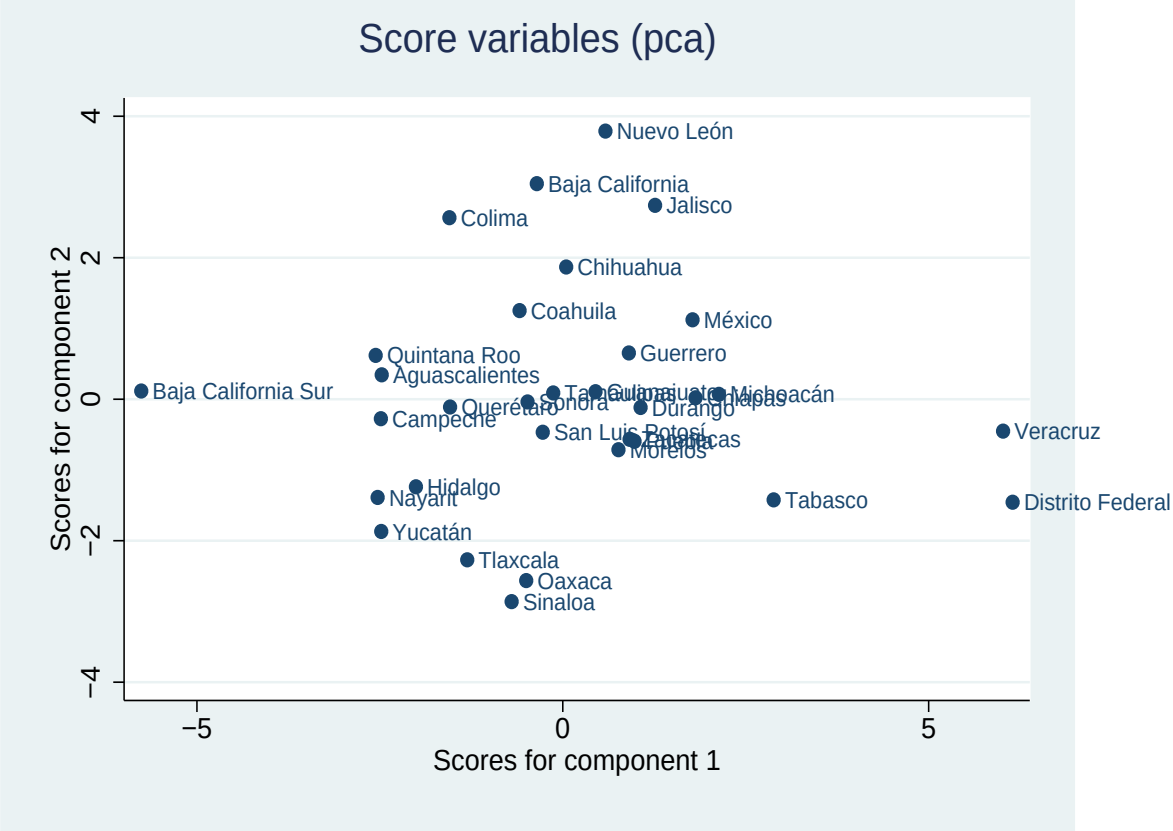
Observaciones:

1. Con esta gráfica de componentes principales se acentúa y se hace evidente el por qué de los resultados obtenidos en **4.3.2.4 y 4.3.2.6** al separarse los estados de Tabasco, Tlaxcala, Veracruz y el Distrito Federal.

2. Se observa la importancia de ambos componentes para conformar las regiones.

3. Se corrobora el comportamiento en cuanto a la "cercanía" de Yucatán, Campeche y Baja California Sur.

4.3.9. Gráfica de los dos primeros componentes principales obtenidos con la matriz de correlaciones.



El porcentaje de varianza explicado por las primeras dos componentes es 68 %.

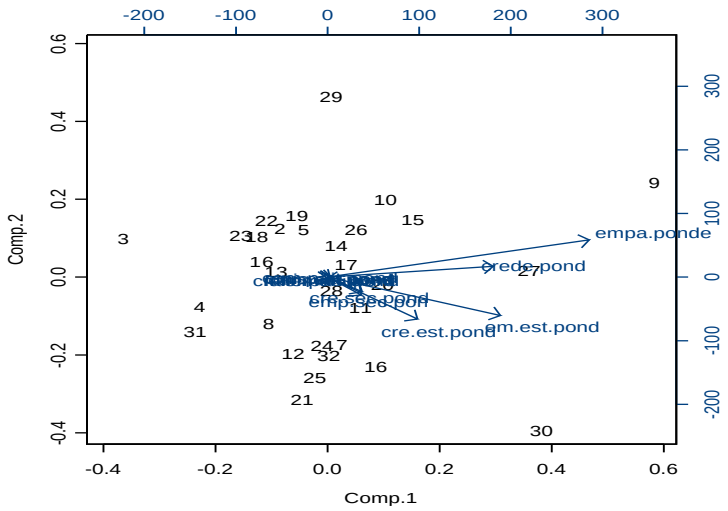
Observaciones:

1. Se confirma el comportamiento de los estados de Tabasco, Veracruz y Distrito Federal, y no tanto así el de Tlaxcala respecto a la gráfica anterior.

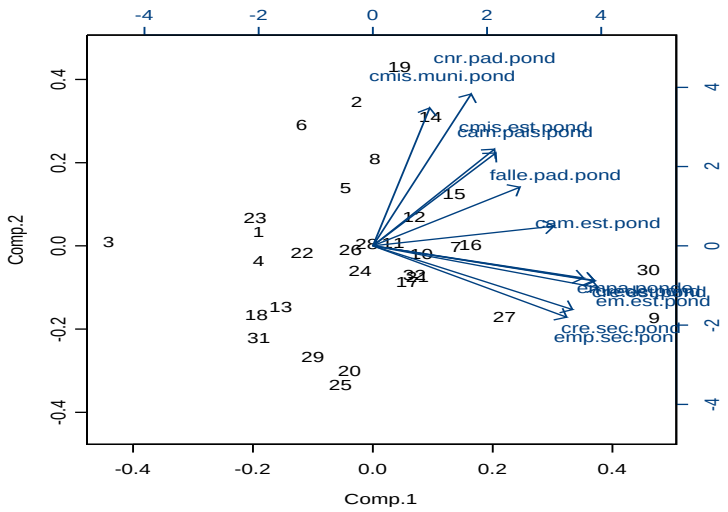
2. En la parte superior se observa el grupo de estados que en parte de los análisis de conglomerados conformaron una región, a decir: Colima, Chihuahua, Coahuila, México, Nuevo León, Jalisco y Baja California.

4.3.10. Gráficas Biplot de los componentes principales obtenidos con los indicadores ponderados.

4.3.10.1. Biplot con la matriz de varianzas y covarianzas.



4.3.10.2. Biplot con la matriz de correlaciones.



Los Biplots son semejantes entre indicadores ponderados en ambas Verificaciones pero no mantienen relación (ni en comportamiento de variables ni en

conformación de regiones) con los resultados previamente obtenidos con los indicadores *en bruto*. Nuevamente Baja California Sur (3) y Distrito Federal (9) a quienes ahora se les une Tlaxcala (29) y Veracruz(30), presentan un comportamiento distinto al resto de las entidades. En ninguna gráfica se presentan de manera clara conjuntos de variables relacionadas como en Biplots anteriores. Sin embargo y particularmente en el Biplot con la matriz de correlaciones, se notan grupos de estados que en análisis previos esta *semejanza o agrupamiento* se había detectado.

Con base en los resultados, es importante mencionar que en este análisis (VNM2006 con indicadores ponderados) las regiones obtenidas no preservan una estructura (ni se marca una tendencia). Nuevamente, se puede ver una notable discrepancia en cuanto a las regiones obtenidas utilizando como medida de disimilaridad la norma euclideana y aquellas obtenidas usando como medida de similaridad la matriz de correlaciones.

Además, estos resultados no presentan *semejanza* alguna con los resultados de la VNM2005 con indicadores ponderados. Esto es contrario a lo ocurrido al analizar los datos en bruto (secciones 4.1 y 4.2), en donde sí se presentaron notables semejanzas tanto en el número de regiones como en las entidades fedarativas que las conformaron. Tan es así, que fue posible exhibir regiones en común. Eso no ocurrió en este caso, y en particular, sobresalió el comportamiento de los estados de México, Yucatán y Tlaxcala, entre otros. Esto no se había detectado y como se podrá ver en las siguientes secciones, no se volverá a presentar.

Sobre esta situación con los indicadores ponderados, se pensó en un principio que el comportamiento de los estados, completamente distinto a la forma en que se venían agrupando, se debía a ciertas inconsistencias distribucionales o de magnitud en los errores estándar (EE). Se procedió entonces a verificar el EE de los indicadores (por estado) para tratar de identificar alguna *anomalía* en ellos,

y sin embargo, no fue así. De hecho, se realizaron pruebas de bondad de ajuste de todos los EE y aunque la gran mayoría de ellas resultaron no significativas al ajuste a una distribución normal, se pudo verificar que no se presentan errores estándar *atípicos*.

Se concluyó que la *ponderación* al multiplicar por el recíproco del EE no es la adecuada para reflejar el comportamiento hipotético de las entidades. Una razón para ello y el hecho de no poder exhibir regiones, es que cada indicador tiene un ponderador diferente en cada estado. Es decir, se tienen ponderadores entre estados y entre indicadores.

De acuerdo con lo mencionado, no es posible exponer regiones que permitan mostrar un tendencia en cuanto a la formación de las mismas, o bien, regiones que apoyen o desaprueben los resultados obtenidos en las secciones 4.1 y 4.2 (con los datos no ponderados).

Sin embargo, pueden surgir preguntas naturales tales como ¿Qué sucede si ahora consideramos en conjunto los indicadores[17] y no por separado[18]?, o bien, ¿Qué ocurre si ponderamos los indicadores (es decir, asignarles un "*peso*") y tomar un combinación lineal de ellos (por ejemplo, el promedio)?, o quizás, considerar ahora combinaciones lineales de los indicadores de la forma:

$$\omega^{i,j} = \alpha \cdot \omega^{i,j}_{2005} + (1-\alpha) \cdot \omega^{i,j}_{2006} \qquad 0 \leq \alpha \leq 1.$$

asignando un mayor peso a los indicadores que, en teoría, presenten una mayor precisión con base en el tamaño de muestra utilizado para llevar a cabo las Verificaciones. Estos cuestionamientos se abordarán y se tratarán de responder en las secciones siguientes.

[17] Es decir, definir nuevos indicadores como combinaciones lineales de ambas Verificaciones.
[18] Como se ha realizado hasta este momento.

4.4. Resultados obtenidos con indicadores de 2005 y 2006.

En esta sección se presentan los resultados obtenidos con los indicadores de ambas verificaciones (VNM2005 y VNM2006) que han sido considerados para este proyecto. Es decir, se tomarán los $2p$ indicadores y las 32 entidades federativas de tal forma que la matriz de datos será de $32 \times 2p$. Hasta ahora se han tratado los indicadores de cada Verificación por separado, por lo cual, los objetivos en este análisis son, por un lado, corroborar resultados que se tienen hasta el momento, y por otro, con el empleo de todos los datos se espera que los resultados reflejen una mayor aproximación a la realidad, dando un mayor peso a aquellos que se consideran más importantes. De igual manera que en las secciones previas [19], en las subsecciones siguentes se exhibirán:

1. Los dendrogramas resultantes del Análisis de Conglomerados y las regiones formadas al proceder con los cortes correpondientes.

2. Las regiones obtenidas con el Algoritmo de Partición k-medias para cuatro y cinco clusters[20]

3. Gráfica de los dos primeros componentes principales obtenidos con la matriz de varianzas y covarianzas.

4. Gráfica de los dos primeros componentes principales obtenidos con la matriz de correlaciones.

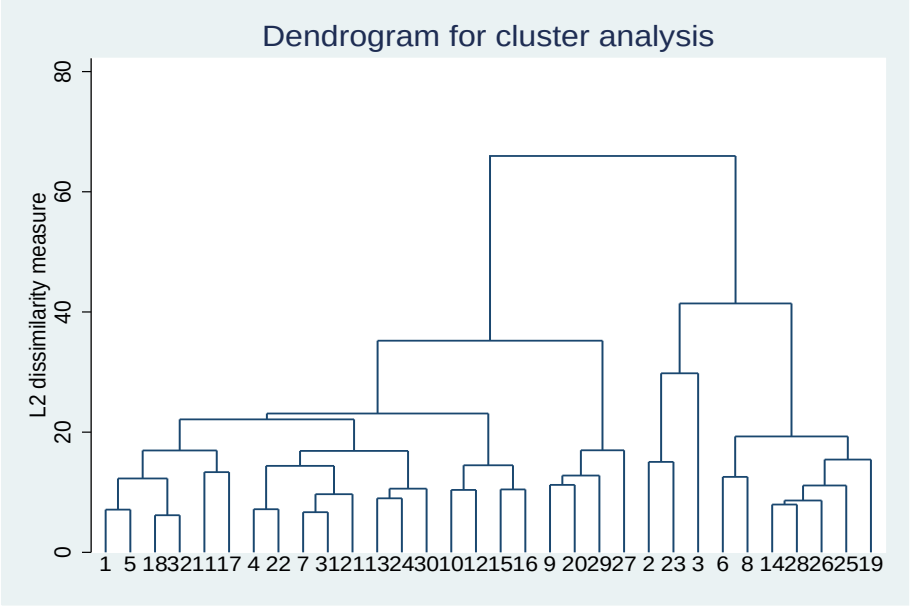
5. Biplots suponiendo homoscedasticidad.

[19] Bajo las mismas hipótesis en los dendrogramas, Algoritmo de K-medias y gráficas de componentes principales, en el sentido de mantener una estructura de regionalización.

[20] Los resultados del Algoritmo de Partición de las k-medias se presentan para cuatro y cinco clusters pues se optó por mantener la estructura obtenida con los dendrogramas.

4.4.1. Análisis de Conglomerados.

4.4.1.1. Análisis de Conglomerados con *liga completa*, tomando como medida de disimilaridad la norma euclideana.



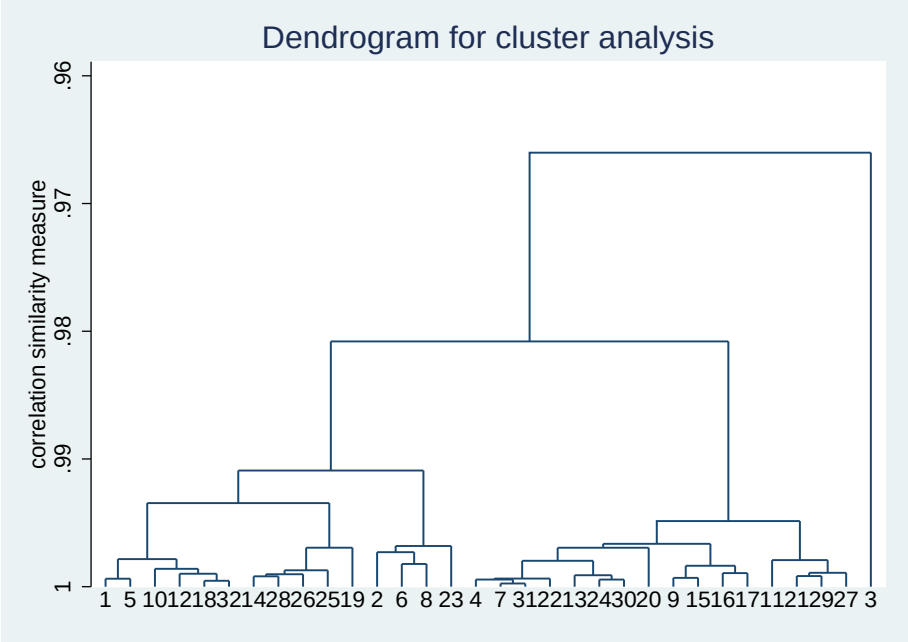
Realizando el corte correspondiente se obtienen las siguientes regiones:

Región 1	(Aguascalientes, Coahuila, Nayarit, Zacatecas, Guanajuato, Morelos), (Campeche, Querétaro, Chiapas, Yucatán, Puebla, Hidalgo, San Luis Potosí, Veracruz), (Durango, Guerrero, México, Michocán).
Región 2	Distrito Federal, Oaxaca, Tlaxcala, Tabasco.
Región 3	Baja California, Quintana Roo, Baja California Sur.
Región 5	Colima, Chihuahua, Jalisco, Tamaulipas, Sonora, Sinaloa, Nuevo León.

Nuevamente, y desde un principio, los estados de Baja California y Quintana Roo se "separan" del resto formando una sola región y aunque Baja California aparece *aislado* en el dendrograma, presenta el mismo comportamiento que estos dos últimos[21].

[21] En ese sentido, puede considerarse que las tres entidades conforman una misma región como se muestra en el cuadro con los paréntesis.

4.4.1.2. Análisis de Conglomerados con *liga completa*, tomando como medida de similaridad la matriz de correlaciones.

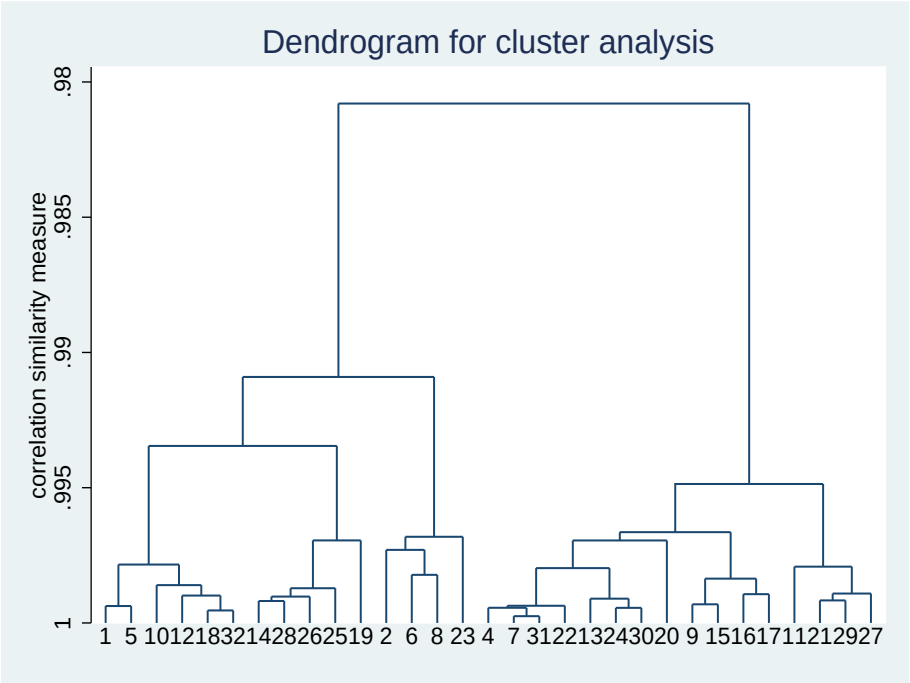


Al realizar el corte correspondiente respecto a la medida de similaridad, se obtienen las siguientes regiones:

Región 1	(Aguascalientes, Coahuila, Durango, Guerrero, Nayarit, Zacatecas), (Jalisco, Tamaulipas, Sonora, Sinaloa, Nuevo León).
Región 2	Baja California, Colima, Chihuahua, Quintana Roo.
Región 3	(Campeche, Chiapas, Yucatán, Querétaro, Hidalgo, San Luis Potosí, Veracruz, Oaxaca, Distrito Federal, México, Michoacán, Morelos), (Guanajuato, Puebla, Tlaxcala, Tabasco).
Región 4	Baja California Sur.

Nótese que, aunque se les unen Colima y Chihuahua (comparando con el caso anterior), los estados de Baja California y Quintana Roo, se mantienen en una región, mientras que Baja California Sur se une al final y nuevamente se "aisla" del resto formando una región. Finalmente, con base en el dendrograma se puede hablar, en función de la altura del corte, de seis regiones como se muestra en la tabla.

Nótese que Baja California Sur se *une* al final en el dendrograma, por lo que podría modificar el dendrograma. Se procedió entonces a eliminarlo con el objetivo de analizar lo que sucede con el dendrograma. Los resultados fueron los siguientes:

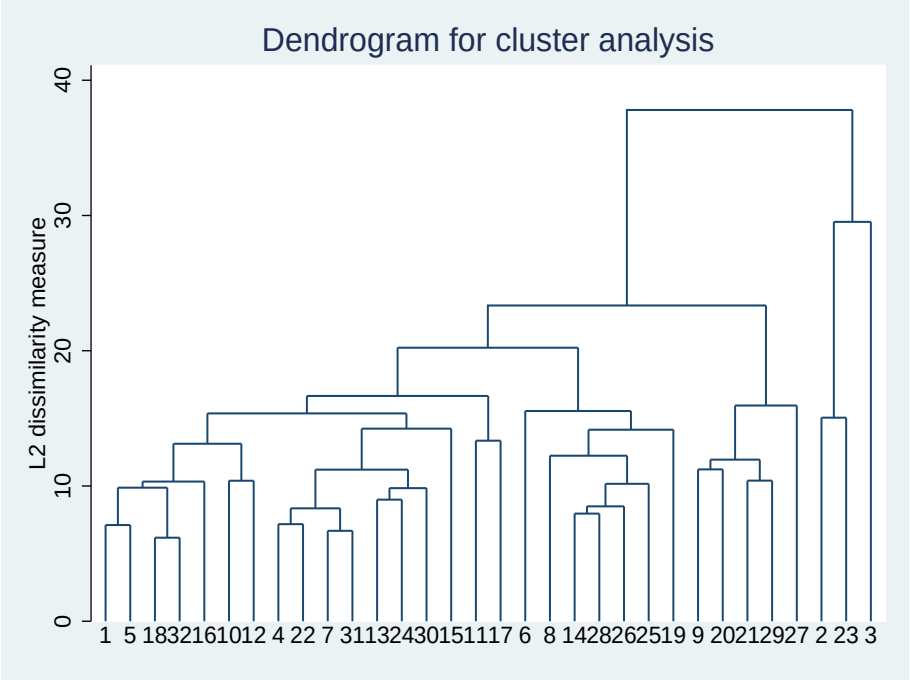


Se observan las mismas regiones que con el dendrograma anterior y se ven más claras las regiones marcadas entre paréntesis:

Región 1	Aguascalientes, Coahuila, Durango, Guerrero, Nayarit, Zacatecas.
Región 2	Jalisco, Tamaulipas, Sonora, Sinaloa, Nuevo León).
Región 3	Baja California, Colima, Chihuahua, Quintana Roo.
Región 4	Campeche, Chiapas, Yucatán, Querétaro, Hidalgo, San Luis Potosí, Veracruz, Oaxaca, Distrito Federal, México, Michoacán, Morelos.
Región 5	Guanajuato, Puebla, Tlaxcala, Tabasco.

Otro punto importante a mencionar es que estas regiones mantienen semejanza con las obtenidas al analizar las Verificaciones de 2005 y 2006 (con datos no ponderados). Y aunque esta estructura no es definitiva pues falta ver el resto de los resultados, ya se marca una tendencia en los mismos.

4.4.1.3. Análisis de Conglomerados con *liga promedio*, tomando como medida de disimilaridad la norma euclideana.



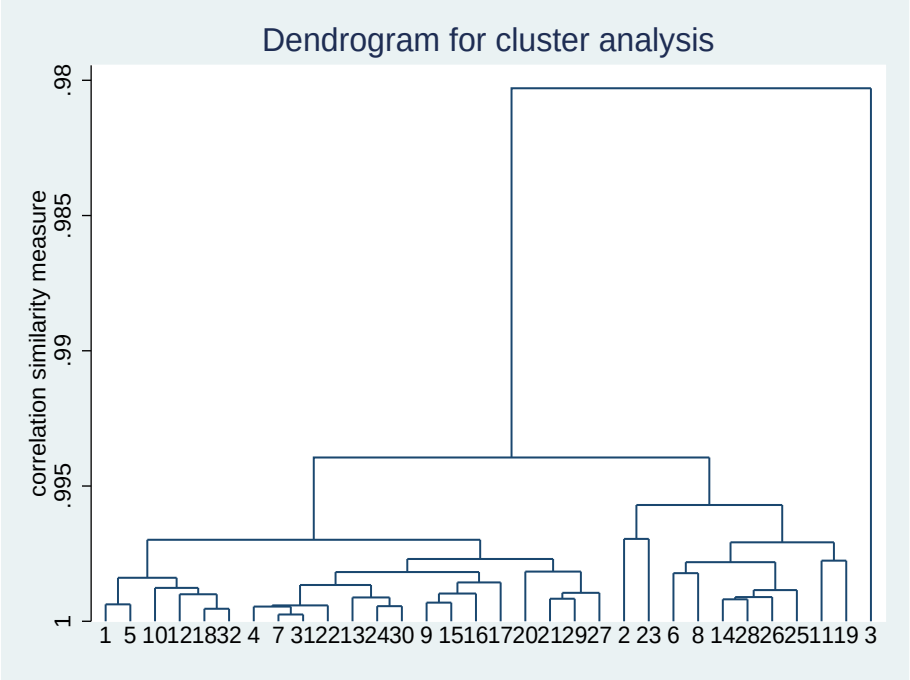
Al ejecutar el corte correspondiente respecto a la medida de disimilaridad, resultan las siguientes regiones:

Región 1	Aguascalientes, Coahuila, Nayarit, Zacatecas, Michoacán, Durango, Guerrero, Campeche, Querétaro, Chiapas, Yucatán, Hidalgo, San Luis Potosí, Veracruz, México, Guanajuato, Morelos.
Región 2	Colima, Chihuahua, Jalisco, Tamaulipas, Sonora, Sinaloa, Nuevo León.
Región 3	Distrito Federal, Oaxaca, Puebla, Tlaxcala, Tabasco.
Región 4	Baja California, Quintana Roo.
Región 5	Baja California Sur.

En este análisis, la región 2 coincide con la región 5 en 4.4.1.1. Algo similar ocurre con la región 3 (de este análisis) y la región 2 de 4.4.1.1.[22]. Sin embargo, en el dendrograma se presenta el efecto de *encadenamiento*, lo cual se ha presentado en gran proporción en los análisis que involucran la *liga promedio* o *peso-promedio ponderado*. Esto complica la identificación de regiones.

[22] Pues D.F., Oaxaca, Tlaxcala y Tabasco, se mantienen en una misma región.

4.4.1.4. Análisis de Conglomerados con *liga promedio*, tomando como medida de similaridad la matriz de correlaciones.

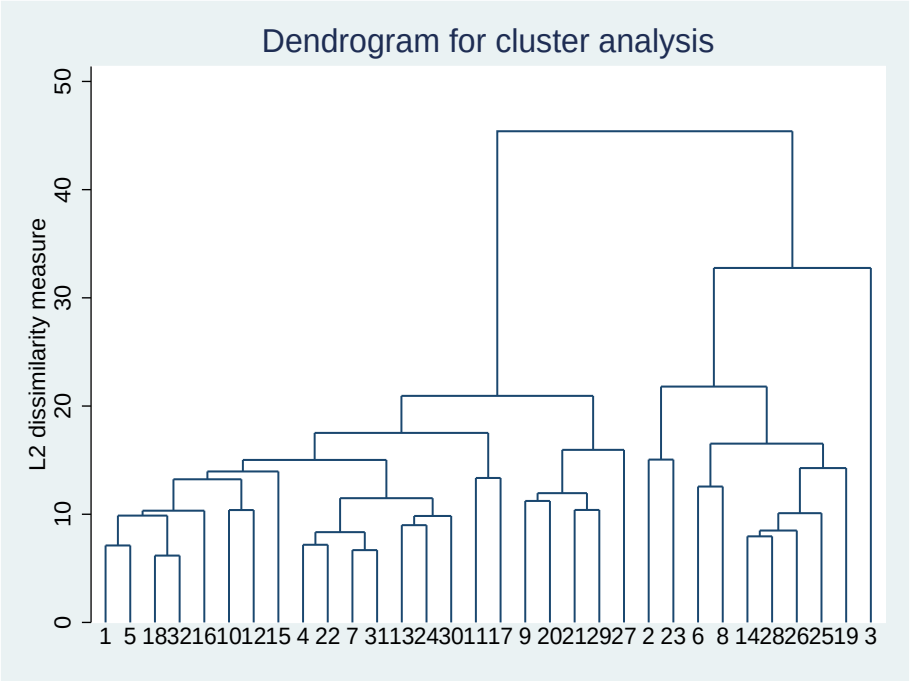


Con el corte correspondiente, se obtienen las siguientes regiones:

Región 1	Aguascalientes, Coahuila, Durango, Guerrero, Nayarit, Zacatecas.
Región 2	Campeche, Chiapas, Yucatán, Querétaro, Hidalgo, San Luis Potosí, Veracruz, Distrito Federal, México, Michoacán, Morelos, Oaxaca, Puebla, Tlaxcala, Tabasco.
Región 3	(Baja California, Quintana Roo.
Región 4	Colima, Chihuahua, Jalisco, Tamaulipas, Sonora, Sinaloa, Guanajuato, Nuevo León).
Región 5	Baja California Sur.

Es claro que las regiones tres y cinco mantienen su estructura respecto a los análisis previos, aunque pueden formar una sola región. Asímismo, la región cuatro esencialmente coincide con las regiones dos de 4.4.1.3 y cinco de 4.4.1.1. lo cual es síntoma de que las estructuras, en esencia, se mantienen.

4.4.1.5. Análisis de Conglomerados con *liga peso - promedio ponderado* (Media de grupos), tomando como medida de disimilaridad la norma euclideana.



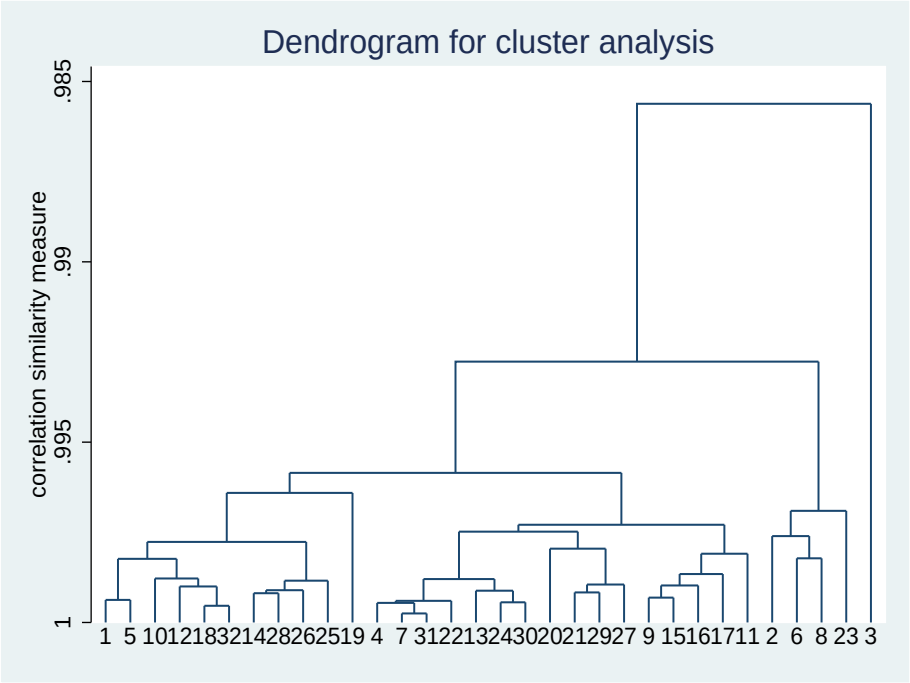
Al realizar el corte correspondiente respecto se obtienen las siguientes regiones:

Región 1	Aguascalientes, Coahuila, Nayarit, Zacatecas, Michoacán, Durango, Guerrero, México, Campeche, Querétaro, Chiapas, Yucatán, Hidalgo, San Luis Potosí, Veracruz, Guanajuato, Morelos.
Región 2	Distrito Federal, Oaxaca, Puebla, Tlaxcala, Tabasco.
Región 3	Baja California, Quintana Roo.
Región 4	Colima, Chihuahua, Jalisco, Tamaulipas, Sonora, Sinaloa, Nuevo León.
Región 5	Baja California Sur.

Si se observan los resultados obtenidos en las subsecciones previas 4.4.1.3 y 4.4.1.1. se notará una consistencia en los estados que conforman las cinco regiones, lo cual indica una *estabilidad* en la estructura de las mismas. Sin embargo, se presenta el efecto de *encadenamiento* en el dendrograma [23] , lo cual complica la identificación de regiones.

[23] Así como en los análisis que involucran la *liga promedio*.

4.4.1.6. Análisis de Conglomerados con *liga peso - promedio ponderado* (Media de grupos), tomando como medida de similaridad la matriz de correlaciones.

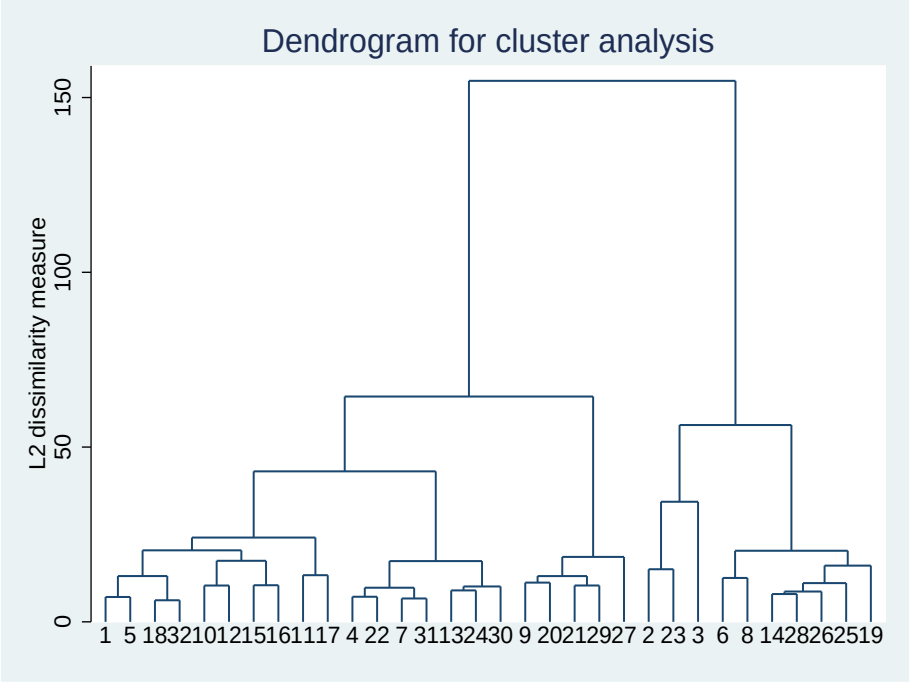


Al ejecutar el corte correspondiente respecto a la medida de similaridad, se obtienen las siguientes regiones:

Región 1	Aguascalientes, Coahuila, Durango, Guerrero, Nayarit, Zacatecas, Jalisco, Tamaulipas, Sonora, Sinaloa, Nuevo León..
Región 2	Campeche, Chiapas, Yucatán, Querétaro, Hidalgo, San Luis Potosí, Veracruz, Oaxaca, Puebla, Tlaxcala, Tabasco, Distrito Federal, México, Michoacán, Morelos, Guanajuato.
Región 3	Baja California, Colima, Chihuahua, Quintana Roo.
Región 4	Baja California Sur.

Se obtienen las mismas regiones que en 4.4.1.2. Además, el comportamiento de Baja California y Quintana Roo se mantiene al permanecer a una misma región; y lo mismo sucede con Baja California Sur.

4.4.1.7. Análisis de Conglomerados con el *Método de Ward*, tomando como medida de disimilaridad la norma euclideana.



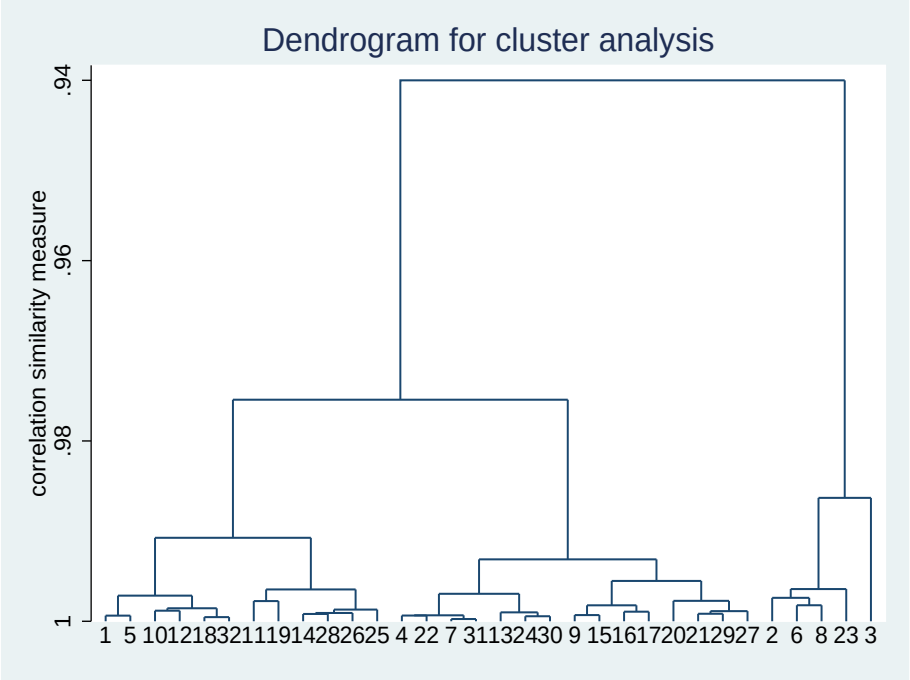
Realizando el corte correspondiente respecto a la medida de disimilaridad se obtienen las siguientes regiones:

Región 1	Aguascalientes, Coahuila, Nayarit, Zacatecas, Durango, Guerrero, México, Guanajuato, Morelos.
Región 2	Campeche, Querétaro, Chiapas, Yucatán, Hidalgo, San Luis Potosí, Veracruz.
Región 3	Distrito Federal, Oaxaca, Puebla, Tlaxcala, Tabasco.
Región 4	Baja California, Quintana Roo, Baja California Sur.
Región 5	Colima, Chihuahua, Jalisco, Tamaulipas, Sonora, Sinaloa, Nuevo León.

En esencia, los estados que conforman las regiones 3, 4 y 5 conservan la misma estructura que en los análisis previos [24]. Nótese del dendrograma que Baja California Sur podría separarse para formar una sola región (en esta ocasión aparece en la misma región que Baja California y Quintana Roo).

[24] Aunque denominados con otro número de región.

4.4.1.8. Análisis de Conglomerados con el *Método de Ward*, tomando como medida de similaridad la matriz de correlaciones.



Al realizar el corte correspondiente respecto a la medida de similaridad se obtienen las siguientes regiones:

Región 1	Aguascalientes, Coahuila, Durango, Guerrero, Nayarit, Zacatecas.
Región 2	Guanajuato, Nuevo León, Jalisco, Tamaulipas, Sonora, Sinaloa.
Región 3	Campeche, Querétaro, Chiapas, Yucatán, Hidalgo, San Luis Potosí, Veracruz.
Región 4	Distrito Federal, México, Michoacán, Morelos, Oaxaca, Puebla, Tlaxcala, Tabasco.
Región 5	Baja California, Colima, Chihuahua, Quintana Roo, Baja California Sur.

Vale la pena mencionar que utilizando el *Método de Ward* se han obtenido los dendrogramas más claros y precisos para formar las regiones. Nótese además que estos resultados coinciden con las regiones obtenidas en 4.4.1.2 y 4.4.1.4, lo cual es indicativo de que se *conserva* una estructura.

Sin embargo, debido a las características del dendrograma, es posible incluso exhibir seis regiones, pues aunque no es tan claro como en ocasiones previas, el estado de Baja California Sur podría conformar una sola región.

Así, las regiones formadas serían:

Región 1	Aguascalientes, Coahuila, Durango, Guerrero, Nayarit, Zacatecas.
Región 2	Guanajuato, Nuevo León, Jalisco, Tamaulipas, Sonora, Sinaloa.
Región 3	Campeche, Querétaro, Chiapas, Yucatán, Hidalgo, San Luis Potosí, Veracruz.
Región 4	Distrito Federal, México, Michoacán, Morelos, Oaxaca, Puebla, Tlaxcala, Tabasco.
Región 5	Baja California, Colima, Chihuahua, Quintana Roo.
Región 6	Baja California Sur.

Con esta "*corrección*", se obtienen prácticamente los mismos resultados que en análisis previos, principalmente con las regiones resultantes de *encadenamiento completo* tomando como medida de similaridad la matriz de correlaciones sin Baja California Sur (4.4.2.1.).

Finalmente, nótese que estos resultados coinciden en gran parte con aquellos obtenidos de las Verificaciones Nacionales Muestrales de 2005 y 2006 (con datos no ponderados). Con ello se apoya la tendencia seguida en esos casos, y de alguna manera se desaprueban las regiones (o por lo menos, no hay evidencia para apoyarlas) obtenidas con los indicadores ponderados.

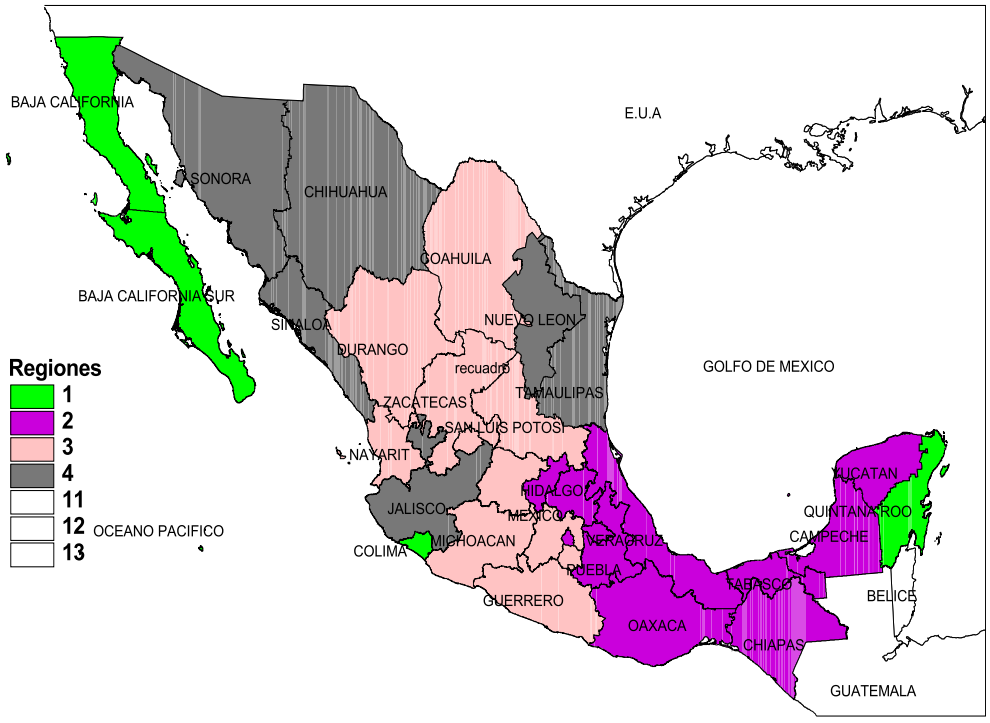
4.4.2. Regiones obtenidas con el Algoritmo de Partición K - medias.

En esta subsección se presentan los resultados obtenidos con el algoritmo de partición k - medias para 4 y 5 regiones.

- **Tomando como medida de disimilaridad la norma euclideana y k observaciones aleatorias como centros iniciales de grupos.**

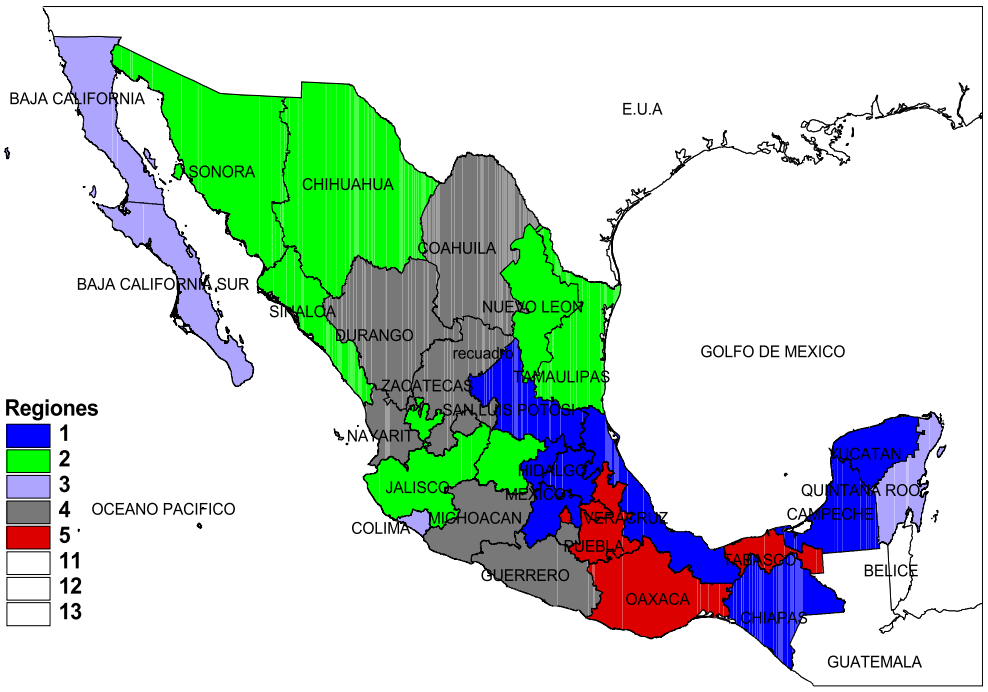
Caso I. k=4 regiones.

Región 1	Aguascalientes, Coahuila, Durango, Guanajuato, Guerrero, México, Michoacán, Morelos, Nayarit, San Luis Potosí, Zacatecas.
Región 2	Campeche, Chiapas, Distrito Federal, Hidalgo, Oaxaca, Puebla, Querétaro, Tabasco, Tlaxcala, Veracruz, Yucatán.
Región 3	Chihuahua, Jalisco, Nuevo León, Sinaloa, Sonora, Tamaulipas.
Región 4	Baja California, Baja California Sur, Colima, Quintana Roo.



Caso II. k=5 regiones.

Región 1	Aguascalientes, Coahuila, Durango, Guerrero, Michocán, Morelos, Nayarit, Zacatecas.
Región 2	Campeche, Chiapas, Hidalgo, México, Querétaro, San Luis Potosí, Veracruz, Yucatán.
Región 3	Chihuahua, Guanajuato, Jalisco, Nuevo León, Sinaloa, Sonora, Tamaulipas.
Región 4	Distrito Federal, Oaxaca, Puebla, Tabasco, Tlaxcala.
Región 5	Baja California, Baja California Sur, Colima, Quintana Roo.

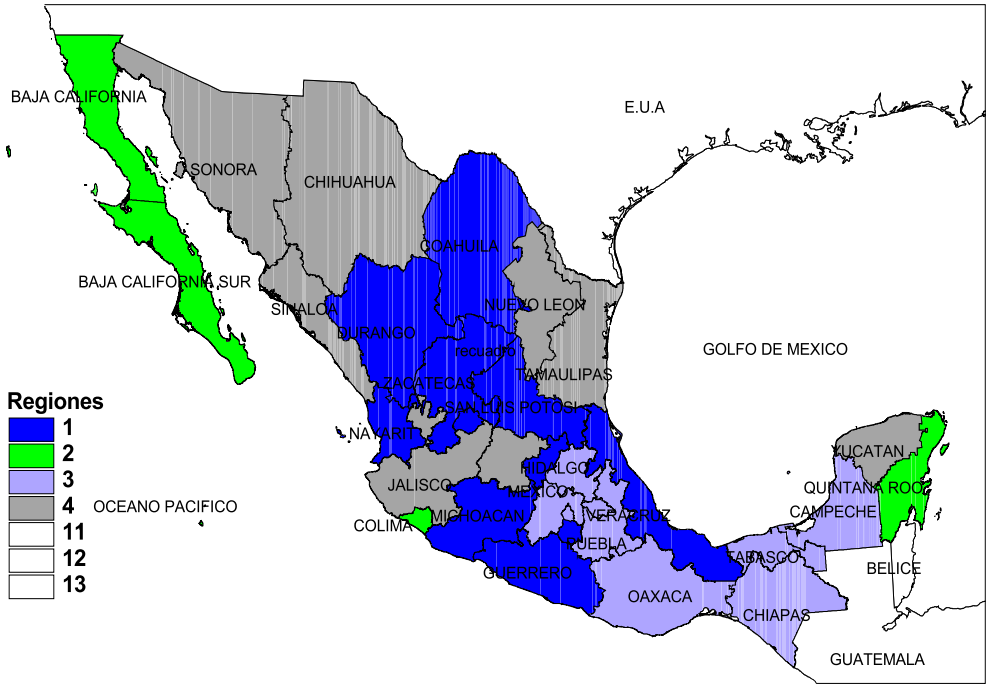


Todas las regiones de ambos casos mantienen una notable semejanza con los resultados del Análisis de Conglomerados y con los de VNM2006. Principalmente, los estados de Oaxaca, Puebla, Tabaco, Tlaxcala y Distrito Federal; lo mismo con los estados que conforman la región 5 previa (BC, BCS, Colima y Quintana Roo).

- **Tomando como medida de similaridad la matriz de correlaciones y k observaciones aleatorias como centros iniciales de grupos.**

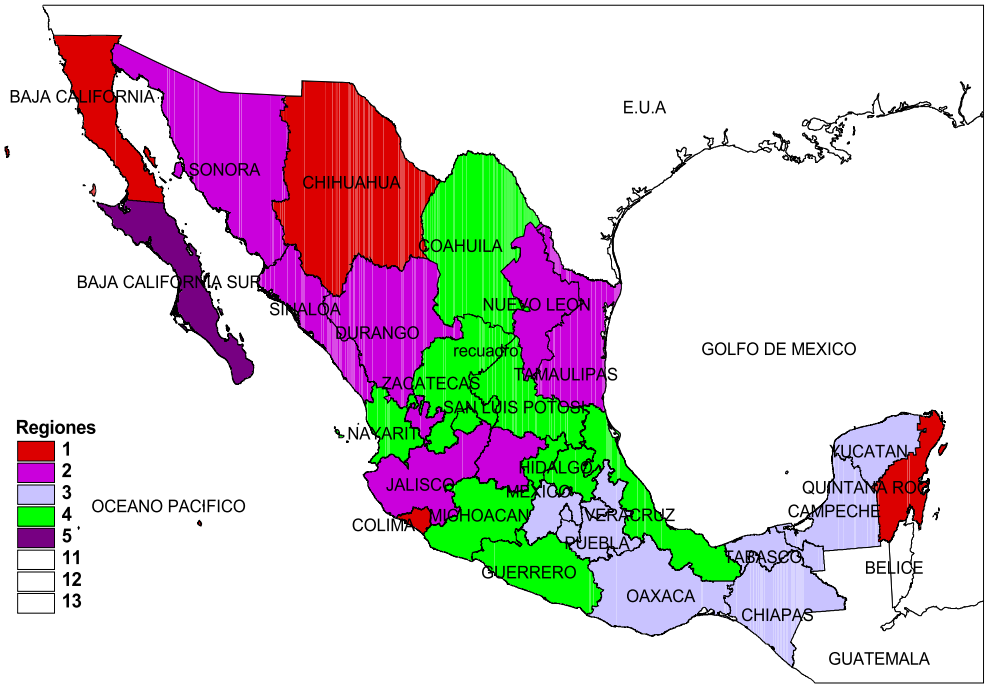
Caso I. k=4 regiones.

Región 1	Aguascalientes, Coahuila, Durango, Guerrero, Michoacán, Morelos, Nayarit, Querétaro, San Luis Potosí, Veracruz, Zacatecas.
Región 2	Campeche, Chiapas, Distrito Federal, Hidalgo, México, Oaxaca, Puebla, Tabasco, Tlaxcala.
Región 3	Yucatán, Chihuahua, Guanajuato, Jalisco, Nuevo León, Sinaloa, Sonora, Tamaulipas.
Región 4	Baja California, Baja California Sur, Colima, Quintana Roo.



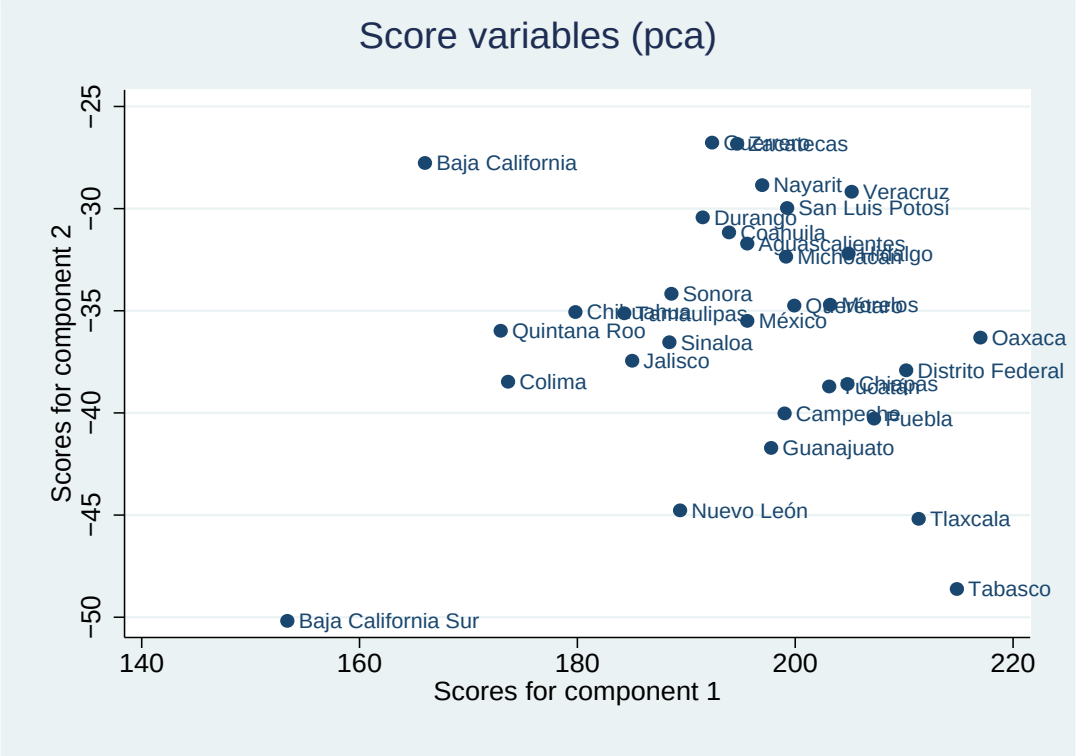
Caso II. k=5 regiones.

Región 1	Aguascalientes, Coahuila, Guerrero, Hidalgo, Michoacán, Nayarit, Querétaro, San Luis Potosí, Veracruz, Zacatecas.
Región 2	Durango, Guanajuato, Jalisco, Nuevo León, Sinaloa, Sonora, Tamaulipas.
Región 3	Campeche, Chiapas, Distrito Federal, México, Morelos, Oaxaca, Puebla, Tabasco, Tlaxcala, Yucatán.
Región 4	Baja California, Chihuahua, Colima, Quintana Roo.
Región 5	Baja California Sur.



Mas allá de recalcar el hecho de que Baja California Sur se aisla y la consistencia de la región formada por Baja California, Chihuahua, Colima y Quintana Roo, es importante notar que hay un grupo de estados del norte de la República que casi no se ha mencionado pero que también se conservan en una misma región: Nuevo León, Sonora, Sinaloa, Jalisco y Tamaulipas; región que también se presentó en los resultados con la VNM20006 ! Esto apoya una *estructura*.

4.4.3. Gráfica de los dos primeros componentes principales obtenidos con la matriz de varianzas y covarianzas.



El porcentaje de varianza explicado por las primeras dos componentes es 87 %.

De la gráfica correspondiente se observa:

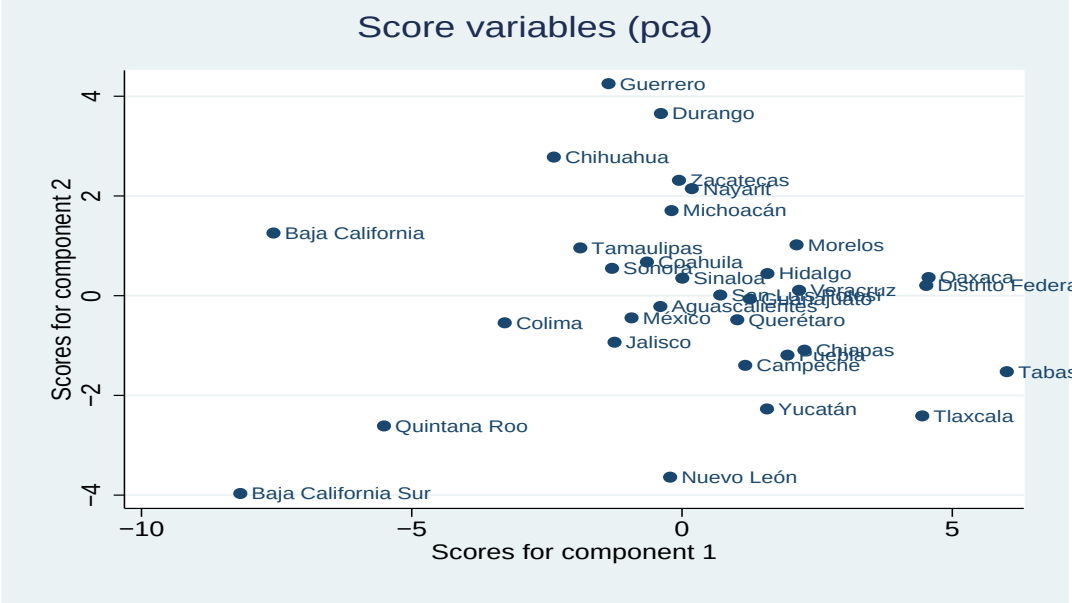
1. El notable "*alejamiento*" de Baja California Sur, lo cual coincide con el comportamiento de ese mismo estado pero para el análisis de la Verificación Nacional 2006 (sin ponderar).

2. La cercanía (respecto al primer componente) entre los estados de Quintana Roo, Colima y Baja California, lo que coincide con los resultados obtenidos en el Análisis de Conglomerados. Y nuevamente, el comportamiento de dichos estados se observa también en el análisis de la Verificación Nacional 2006 (sin ponderar).

3. En el centro de la gráfica se tiene un grupo de estados que, en esencia, mantuvieron su comportamiento en los análisis previos al permanecer a la misma región, a decir: Chihuahua, Sonora, Jalisco, Sinaloa, Tamaulipas, Nuevo León y Querétaro.

4. En la zona inferior-derecha otro grupo de estados que también se asociaron en una misma región: Distrito Federal, Oaxaca, Puebla, Tlaxcala y Tabasco. Para ambas observaciones (tres y cuatro), los comportamientos descritos se presentaron de igual manera que en la VNM2006 (sin ponderar).

4.4.4. Gráfica de los dos primeros componentes principales obtenidos con la matriz de correlaciones.



El porcentaje de varianza explicado por las primeras dos componentes es 67 %.

Nótese que:

1. Se confirma el comportamiento del estado de Baja California Sur por un lado, pues es notable el alejamiento que guarda respecto al complemento de

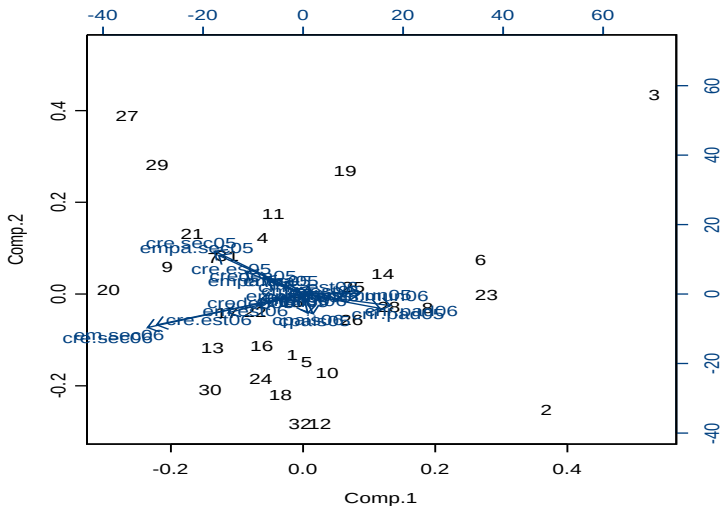
estados. Por otro lado, Baja California y Quintana Roo, a pesar de no estar tan cercanos, se corrobora su "*alejamiento*".

2. Nuevamente y respecto al análisis previo (4.4.3), en la zona inferior-derecha se observa un grupo de estados que estuvieron conformando una misma región: Distrito Federal, Oaxaca, Puebla, Tlaxcala y Tabasco (situación que también se presentó en los análisis correspondientes para las VNM2005 y VNM2006).

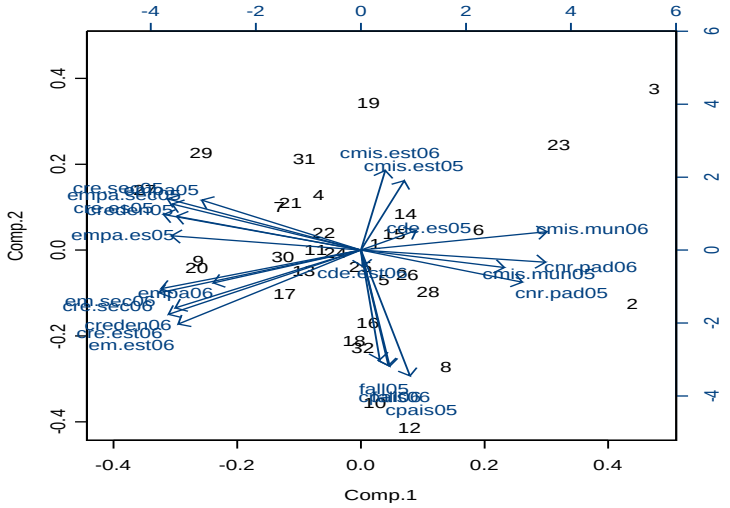
3. En la parte central de la gráfica, se tiene una concentración de los estados que conformaron, en la mayoría de los análisis previos, la **Región 1** o la **Región 2**.

4.4.5. Gráficas Biplot de los componentes principales obtenidos con todos los indicadores.

4.4.5.1. Biplot con la matriz de varianzas y covarianzas.



4.4.5.2. Biplot con la matriz de correlaciones.



Las variables tienen la misma notación que en las secciones previas. Para la VNM2006, se tiene[25]:

Notación	**Variables (representadas con los vectores.)**
empa06	Empadronados en 2006.
empa.es06	Empadronados en el estado en 2006 .
empa.sec06	Empadronados en la sección en 2006.
creden06	Credencializados en 2006.
creden.es06	Credencializados en el estado en 2006.
creden.sec06	Credencializados en la sección en 2006.
falle06	Fallecidos en 2006.
cnr.pad06	Cambios de domicilio no reportados en el padrón en 2006.
cmis.mu06	Cambios de domicilio al mismo municipio en 2006.

[25] De manera análoga se obtienen los nombres de las variables para la VNM2005, cuya tabla puede verse en la sección 4.1.5.

Notación	**Variables (representadas con los vectores.)**
cmis.es06	Cambios de domicilio a otro municipio dentro del mismo estado en 2006.
ced.es06	Cambios de domicilio a otro estado en 2006.
c.pais06	Cambios de domicilio a otro país en 2006.

En ambas gráficas Biplots se observan, el *alejamiento* de Baja California Sur, y nuevamente la estructura de Baja California (2), Quintana Roo (23) y Colima (6) al *alejarse* del resto. Incluso, en los resultados de esta sección se comentó que Baja California Sur podría unirse a la región conformada por Baja California, Quintana Roo y Colima, lo cual se corrobora con los Biplots. Respecto a las variables, no es claro el comportamiento para la primer gráfica, pero en la segunda se observa que los indicadores correspondientes para cada Verificación (2005 y 2006) mantienen una estructura de correlación, indicativo de consistencia en las variables. Por ejemplo, entre los grupos de indicadores *Cambios de domicilio al mismo municipio 2005 y 2006*, y *Cambios no reportados al Padrón 2005 y 2006*.

Como observaciones generales para esta sección, nótese que se tiene un grupo de estados (del norte): Sonora, Sinaloa, Tamaulipas, Chihuahua y Nuevo León, que pertenecen en cada análisis a la misma región. Esto indica un comportamiento similar en los estados del norte que es presentado también por estados del centro. Asimismo las estimaciones de la VNM2006 están teniendo un mayor "*peso*" pues el comportamiento de los estados para la formación de las regiones (tanto para Análisis de Conglomerados, el Algoritmo de k-medias y Componentes Principales) es similar a los correspondientes análisis llevados a cabo para dicha Verificación. Esto puede interpretarse como un comportamiento lógico pues la VNM2006 es la más reciente.

Finalmente, aunque las regiones formadas durante los análisis realizados hasta ahora no son constantes en su totalidad (pero sí lo son en la mayoría de ellos), nuevamente se observa que la *estructura* en el comportamiento de las entidades no coincide con la regionalización propuesta por CONAPO. Esto no indica que dicha regionalización sea incorrecta, sino que el comportamiento electoral en el país no coincide con la distribución de la marginación.

4.5. Resultados con el promedio aritmético de los indicadores.

A continuación, se exhiben los resultados obtenidos con "*nuevas variables*" definidas como el promedio aritmético de los indicadores en ambas Verificaciones. Es decir, si denotamos como $\omega^{i,j}_{2005}$ y $\omega^{i,j}_{2006}$ los i-ésimos indicadores correspondientes a la VNM2005 y VNM2006 respectivamente para el j-ésimo estado, entonces:

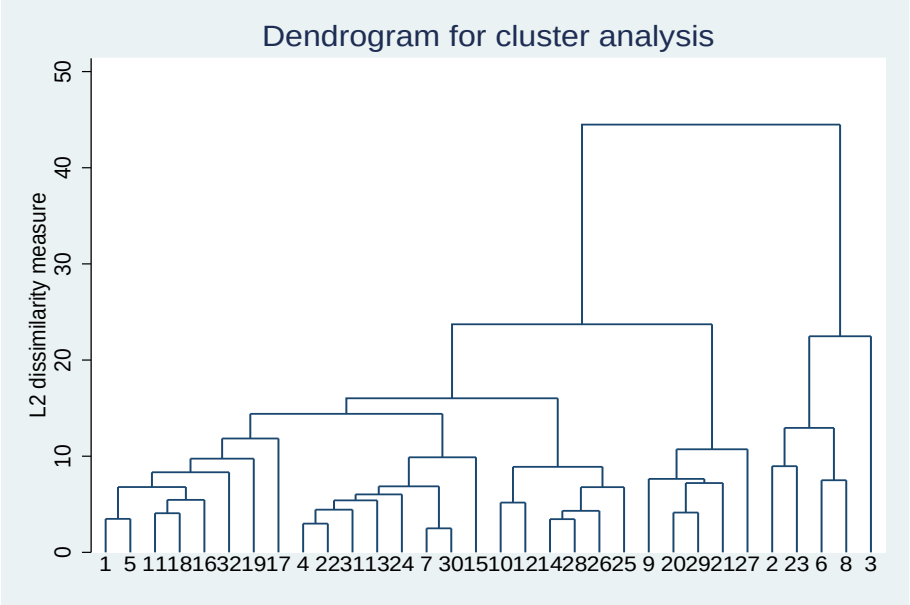
$$\omega^{i,j} = \frac{\omega^{i,j}_{2005} + \omega^{i,j}_{2006}}{2}$$

es el "*nuevo indicador*"definido para la j - ésima entidad federativa. El objetivo es verificar el comportamiento de los estados y comparar con las regiones obtenidas. Los resultados se exhibirán al mismo tenor que en las secciones previas.

En los resultados de esta sección se podrá observar que todos los Análisis de Conglomerados arrojaron cinco regiones; aunque como se ha mencionado, el número de regiones obtenidas se encuentra en función de la altura del corte al dendrograma. En este sentido, es importante mencionar que los resultados con el algoritmo de partición de las k - medias se exhibirán para **cuatro** y **cinco** regiones con el objetivo de compararlos con aquellos que se obtuvieron en las secciones previas (VNM2005, VNM2006, etc.)

4.5.1. Análisis de Conglomerados.

4.5.1.1. Análisis de Conglomerados con *liga completa*, tomando como medida de disimilaridad la norma euclideana.

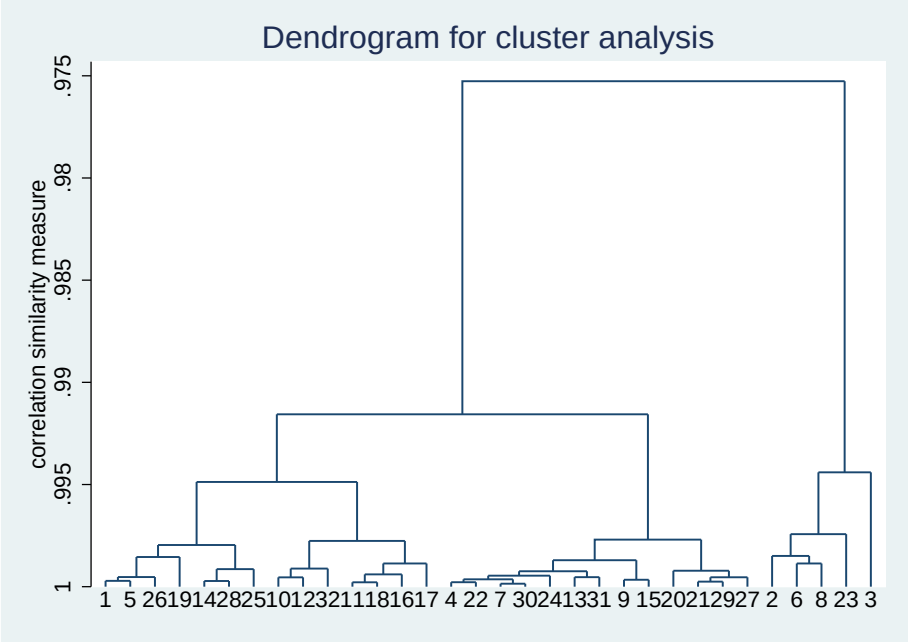


Realizando el corte correspondiente se obtienen las siguientes regiones:

Región 1	(Aguascalientes, Coahuila, Guanajuato, Nayarit, Michoacán, Zacatecas, Nuevo León, Morelos), (Campeche, Querétaro, Yucatán, Hidalgo, San Luis Potosí, Chiapas, Veracruz, México).
Región 2	Durango, Guerrero, Jalisco, Tamaulipas, Sonora, Sinaloa.
Región 3	Distrito Federal, Oaxaca, Tlaxcala, Puebla, Tabasco.
Región 4	(Baja California Sur, Quintana Roo, Colima, Chihuahua.
Región 5	Baja California).

Nótese que todas las regiones (en particular las regiones 1, 3, 4 y 5) mantienen su estructura respecto al análisis de la sección previa y con los resultados obtenidos para la VNM2006. Se presenta el efecto de *encadenamiento* en la formación de la región 1, la cual podría separarse en dos regiones. Mientras que las regiones 4 y 5 se podrían conformar una sola, como lo muestran los paréntesis.

4.5.1.2. Análisis de Conglomerados con *liga completa*, tomando como medida de similaridad la matriz de correlaciones.

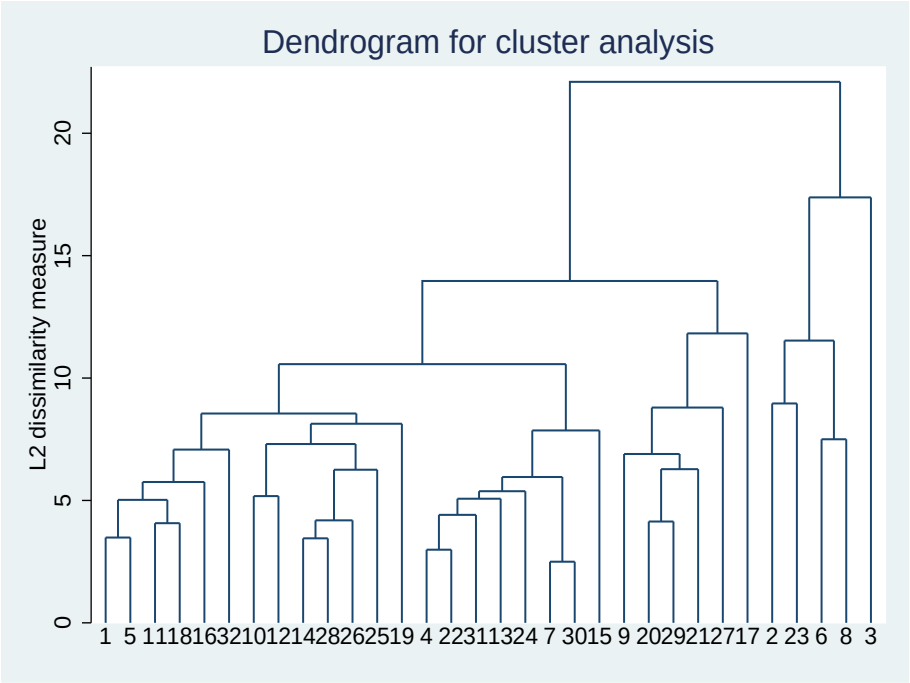


Al realizar el corte correspondiente respecto a la medida de similaridad, se obtienen las siguientes regiones:

Región 1	(Aguascalientes, Coahuila, Sonora, Nuevo León, Jalisco, Tamaulipas, Sinaloa), (Durango, Guerrero, Zacatecas, Guanajuato, Nayarit, Michoacán, Morelos).
Región 2	(Campeche, Querétaro, Chiapas, Veracruz, San Luis Potosí, Hidalgo, Yucatán, Distrito Federal, México.
Región 3	Oaxaca, Puebla, Tlaxcala, Tabasco).
Región 4	Baja California, Colima, Chihuahua, Quintana Roo.
Región 5	Baja California Sur.

El comportamiento de Baja California Sur prevalece, así como la estructura de las regiones 1, 3 y 4 en comparación con los resultados de 4.5.1.1 y las regiones 2, 4 y 5 en comparación con el correspondiente análisis en la VNM2006. Sin embargo, basados en el dendrograma, la región 1 podría separarse mientras que las regiones 2 y 3 se unirían para formar una misma.

4.5.1.3. Análisis de Conglomerados con *liga promedio*, tomando como medida de disimilaridad la norma euclideana.

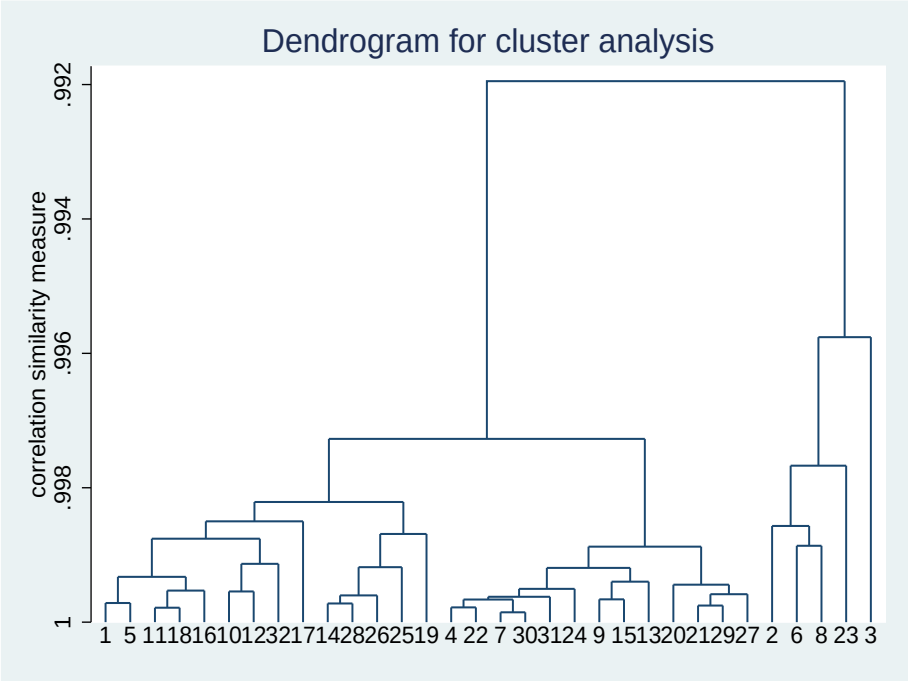


Al ejecutar el corte correspondiente respecto a la medida de disimilaridad, resultan las siguientes regiones:

Región 1	Aguascalientes, Coahuila, Guanajuato, Nayarit, Michocaán, Zacatecas, Durango, Guerrero, Jalisco, Tamaulipas, Sonora, Sinaloa, Nuevo León.
Región 2	Campeche, Querétaro, Yucatán, Hidalgo, San Luis Potosí, Chiapas, Veracruz, México.
Región 3	Distrito Federal, Oaxaca, Tlaxcala, Puebla, Tabasco, Morelos.
Región 4	Baja California, Quintana Roo, Colima, Chihuahua.
Región 5	Baja California Sur.

De nuevo se presenta un efecto de *encadenamiento*, lo cual dificulta la determinación de las regiones. Sin embargo, nótese que la región 2 de 4.5.1.1 forma parte de la región 1 para este caso (Durango, Guerrero,..., Sinaloa). Mientras que la región 2 de este análisis, forma parte de la región 1 en 4.5.1.1. Las regiones 3, 4 y 5 se conservan.

4.5.1.4. Análisis de Conglomerados con *liga promedio*, tomando como medida de similaridad la matriz de correlaciones.

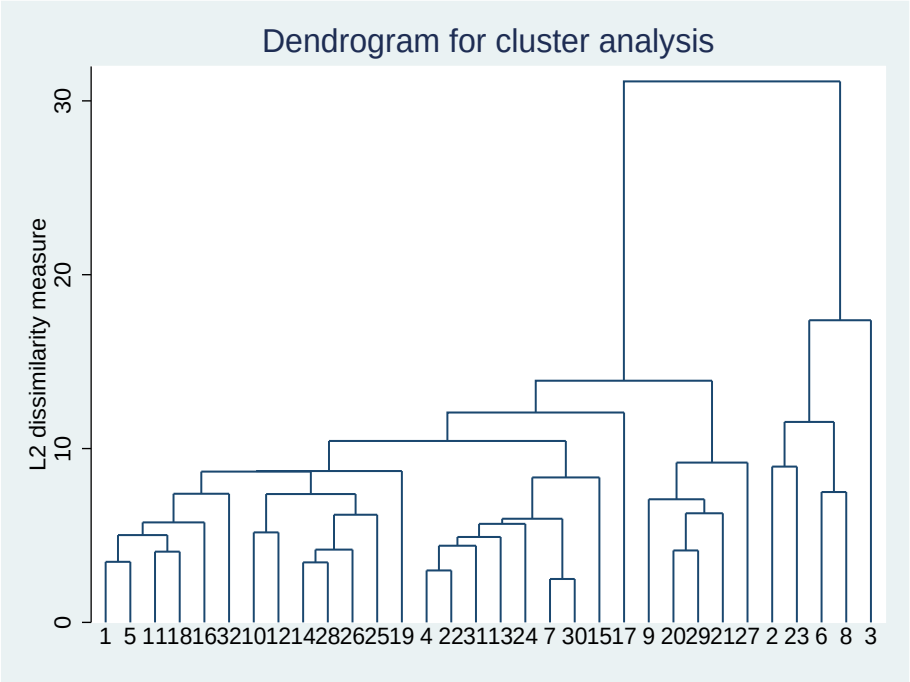


Con el corte correspondiente, se obtienen las siguientes regiones:

Región 1	Aguascalientes, Coahuila, Guanajuato, Nayarit, Michoacán, Durango, Guerrero, Zacatecas, Morelos.
Región 2	Jalisco, Tamaulipas, Sonora, Sinaloa, Nuevo León.
Región 3	Campeche, Querétaro, Chiapas, Veracruz, Yucatán, San Luis Potosí, Distrito Federal, México, Hidalgo, Oaxaca, Puebla, Tlaxcala, Tabasco.
Región 4	Baja California, Colima, Chihuahua, Quintana Roo.
Región 5	Baja California Sur.

A excepción de la región dos, las restantes comparten esencialmente la misma estructura con las regiones de los análisis previos. Sin embargo, y debido a la cercanía en el dendrograma, se puede considerar que las regiones 1 y 2 para este caso, forman una misma.

4.5.1.5. Análisis de Conglomerados con *liga peso - promedio ponderado* (Media de grupos), tomando como medida de disimilaridad la norma euclideana.

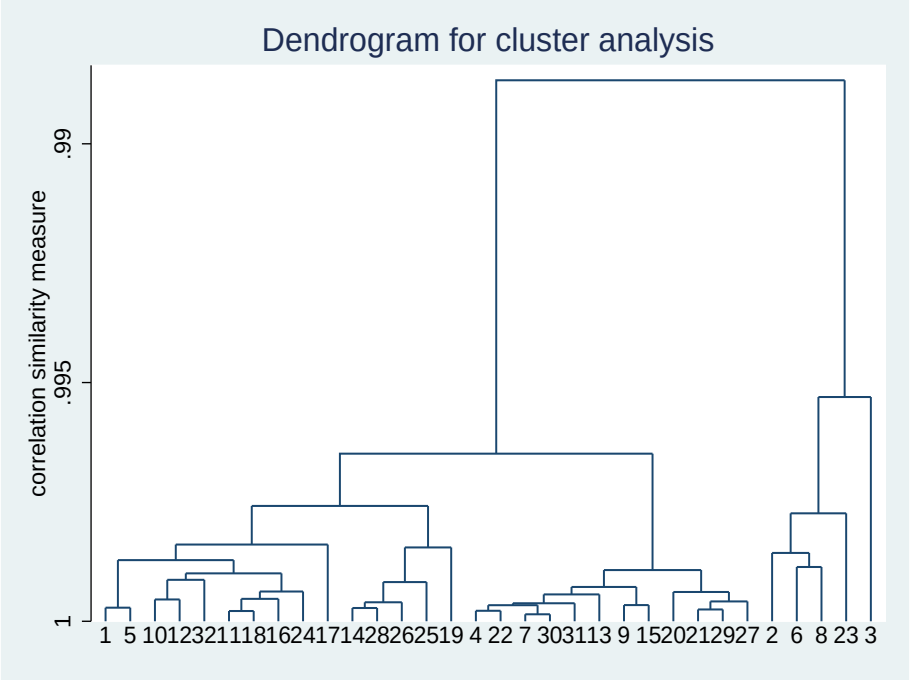


Al realizar el corte correspondiente se obtienen las siguientes regiones:

Región 1	Aguascalientes, Coahuila, Guanajuato, Nayarit, Michoacán, Zacatecas, Durango, Guerrero, Jalisco, Tamaulipas, Sonora, Sinaloa, Nuevo León.
Región 2	Campeche, Querétaro, Yucatán, Hidalgo, San Luis Potosí, Chiapas, Veracruz, Meźico.
Región 3	Distrito Federal, Oaxaca, Tlaxcala, Puebla, Tabasco, Morelos.
Región 4	Baja California, Quintana Roo, Colima, Chihuahua.
Región 5	Baja California Sur.

El efecto de *encadenamiento* se presenta de nuevo y ha sido común al utilizar la *liga peso - promedio* y la norma euclideana. El estado cuyo comportamiento llama la atención en este análisis es Morelos (17) pues en el dendrograma se encuentra más cercano hacia la región 2. Las restantes regiones se mantienen.

4.5.1.6. Análisis de Conglomerados con *liga peso - promedio ponderado* (Media de grupos), tomando como medida de similaridad la matriz de correlaciones.

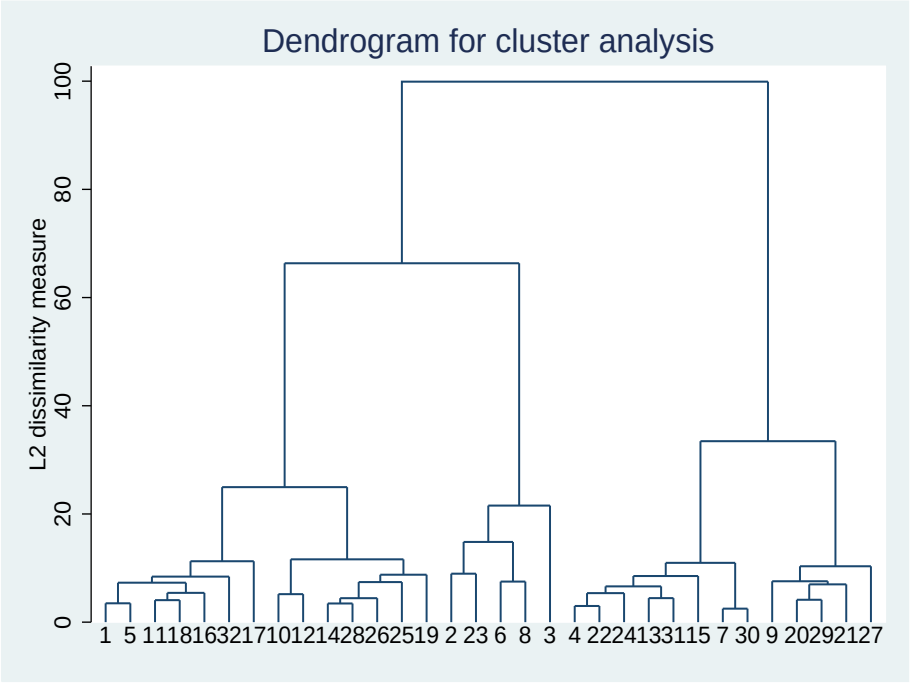


Al ejecutar el corte correspondiente respecto a la medida de similaridad, se obtienen las siguientes regiones:

Región 1	Aguascalientes, Coahuila, Durango, Guerrero, Zacatecas, Guanajuato, Nayarit, Michoacán, San Luis Potosí, Morelos.
Región 2	Jalisco, Tamaulipas, Sonora, Sinaloa, Nuevo León.
Región 3	Campeche, Querétaro, Chiapas, Veracruz, Yucatán, Hidalgo, Distrito Federal, México, Oaxaca, Puebla, Tlaxcala, Tabasco.
Región 4	Baja California, Colima, Chihuahua, Quintana Roo.
Región 5	Baja California Sur.

Esencialmente, se obtienen las mismas regiones que en 4.5.1.4. Además, el comportamiento de Baja California y Quintana Roo se mantiene al permanecer a una misma región; y lo mismo sucede con Baja California Sur, lo cual indica una consistencia interna con las regiones resultantes.

4.5.1.7. Análisis de Conglomerados con el *Método de Ward*, tomando como medida de disimilaridad la norma euclideana.

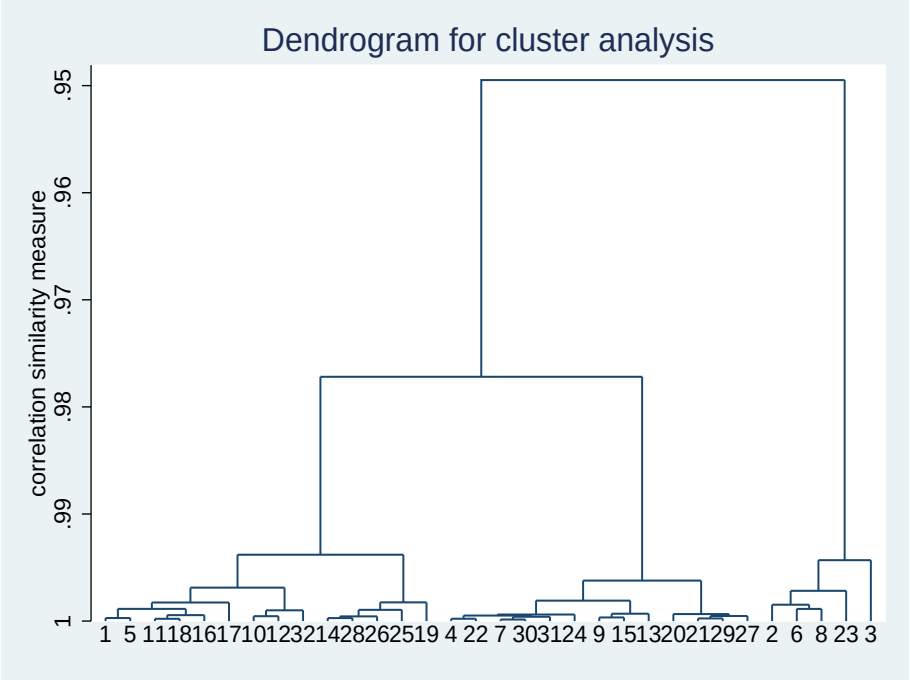


Realizando el corte correspondiente respecto a la medida de disimilaridad se obtienen las siguientes regiones:

Región 1	Aguascalientes, Coahuila, Guanajuato, Nayarit, Michoacán, Zacatecas, Morelos.
Región 2	Durango, Guerrero, Jalisco, Tamaulipas, Sonora, Sinaloa, Nuevo León.
Región 3	Baja California, Quintana Roo, Colima, Chihuahua, Baja California Sur.
Región 4	Campeche, Querétaro, San Luis Potosí, Hidalgo, Yucatán, México, Chiapas, Veracruz.
Región 5	Distrito Federal, Oaxaca, Tlaxcala, Puebla, Tabasco.

Las regiones 3, 4 y 5 conservan su estructura respecto a los análisis previos. Se tomó al estado de Baja California Sur (3) como parte de la región 3, sin embargo, como puede verse en el dendrograma, podría considerársele en una sola región.

4.5.1.8. Análisis de Conglomerados con el *Método de Ward*, tomando como medida de similaridad la matriz de correlaciones.



Al realizar el corte correspondiente respecto a la medida de similaridad se obtienen las siguientes regiones:

Región 1	Aguascalientes, Coahuila, Guanajuato, Nayarit, Michoacán, Morelos, Durango, Guerrero, Zacatecas.
Región 2	Jalisco, Tamaulipas, Sonora, Sinaloa, Nuevo León.
Región 3	Campeche, Querétaro, Chiapas, Veracruz, Yucatán, Hidalgo, San Luis Potosí, Distrito Federal, México, Hidalgo, Oaxaca, Puebla, Tlaxcala, Tabasco.
Región 4	(Baja California, Colima, Chihuahua, Quintana Roo.
Región 5	Baja California Sur).

Es importante observar que las regiones han mantenido una consistencia interna en cuanto a los estados que las conforman. Esto puede verse con la región dos que contiene estados principalmente del norte de la República Mexicana. Sin embargo, con base en el dendrograma, se podrían considerar a Baja California Sur como parte de la región 4, de tal forma que se tendrían cuatro regiones.

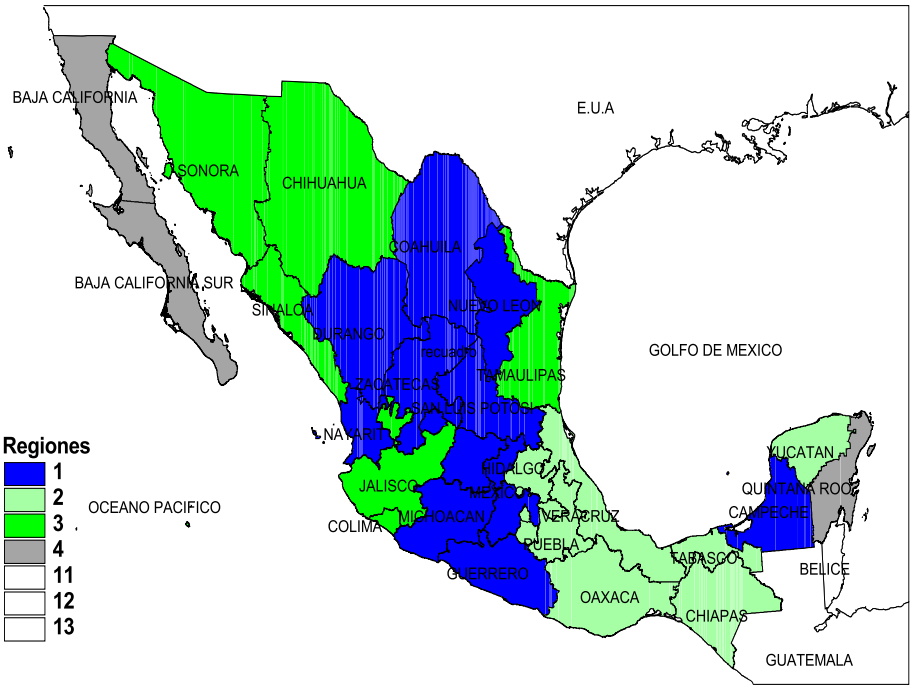
4.5.2. Regiones obtenidas con el Algoritmo de Partición K - medias.

Enseguida se presentan los resultados obtenidos con el algoritmo de partición k - medias para 4 y 5 regiones[26].

- **Tomando como medida de disimilaridad la norma euclideana y k observaciones aleatorias como centros iniciales de grupos.**

Caso I. k=4 regiones.

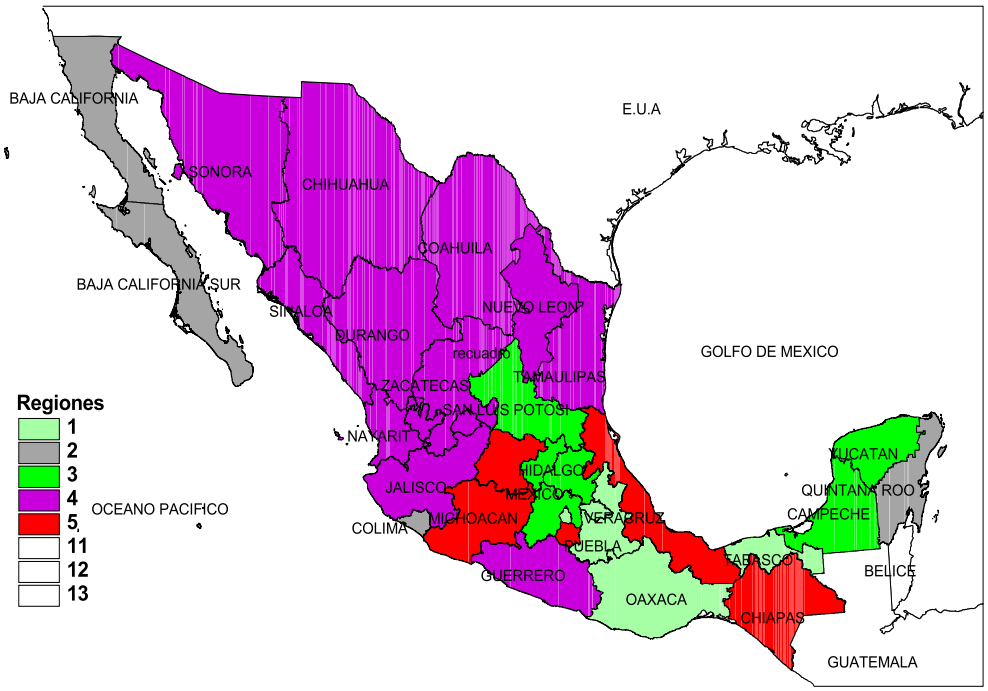
Región 1	Aguascalientes, Campeche, Coahuila, Durango, Guanajuato, Guerrero, México, Michoacán, Nayarit, Nuevo León, Querétaro, San Luis Potosí, Zacatecas.
Región 2	Chiapas, Distrito Federal, Hidalgo, Morelos, Oaxaca, Puebla, Tabasco, Tlaxacala, Veracruz, Yucatán.
Región 3	Colima, Chihuahua, Jalisco, Sinaloa, Sonora, Tamaulipas.
Región 4	Baja California, Baja California Sur, Quintana Roo.



[26] Recuérdese que en el Análisis de Conglomerados se obtuvieron siempre cinco regiones, sin embargo, se exihibirá el caso para cuatro regiones con la intención de comparar resultados.

Caso II. k=5 regiones.

Región 1	Aguascalientes, Nayarit, Chihuahua, Durango, Guerrero, Sinaloa, Sonora, Nuevo León, Jalisco, Coahuila, Tamaulipas, Zacatecas.
Región 2	Campeche, Hidalgo, México, Querétaro, San Luis Potosí, Yucatán.
Región 3	Distrito Federal, Oaxaca, Puebla, Tabasco, Tlaxcala.
Región 4	Chiapas, Guanajuato, Michoacán, Morelos, Veracruz.
Región 5	Baja California, Quintana Roo, Baja California Sur, Colima.

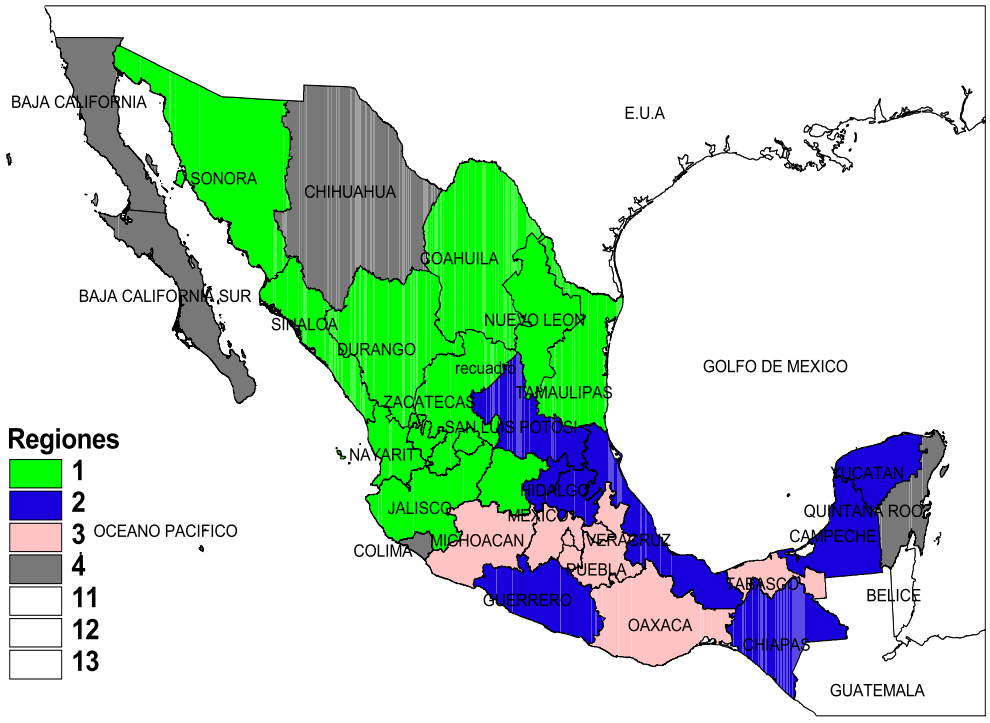


Es claro que la consistencia en las regiones se preserva al pasar de los Análisis de Conglomerados al Algoritmo de las k-medias. Asimismo, al comparar estos resultados con los obtenidos con la VNM2006, se tienen regiones muy similares en cuanto a su consistencia interna. Para el último caso reportado, las regiones 1, 2, 3 y 5 son un buen ejemplo.

- **Tomando como medida de similaridad la matriz de correlaciones y k observaciones aleatorias como centros iniciales de grupos.**

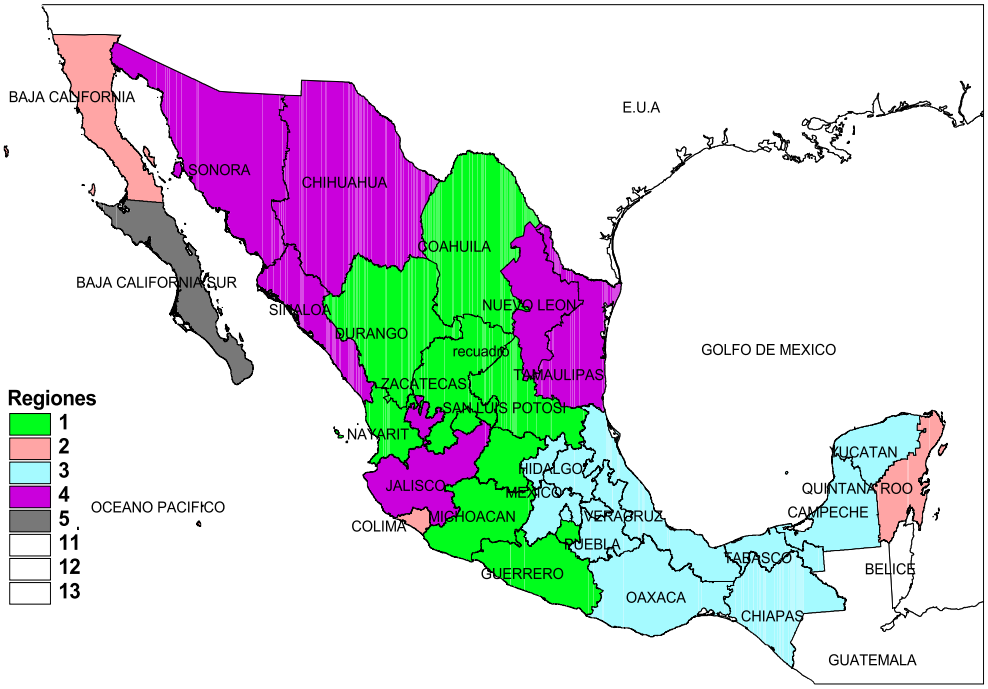
Caso I. k=4 regiones.

Región 1	Aguascalientes, Coahuila, Durango, Guanajuato, Jalisco, Nayarit, Nuevo Léon, Sinaloa, Sonora, Tamaulipas, Zacatecas.
Región 2	Campeche, Chiapas, Guerrero, Hidalgo, Querétaro, Veracruz, San Luis Poosí, Yucatán.
Región 3	Distrito Federal, México, Michoacán, Morelos, Oaxaca, Puebla, Tabasco, Tlaxcala.
Región 4	Baja California, Baja California Sur, Colima, Chihuahua, Quintana Roo.



Caso II. k=5 regiones.

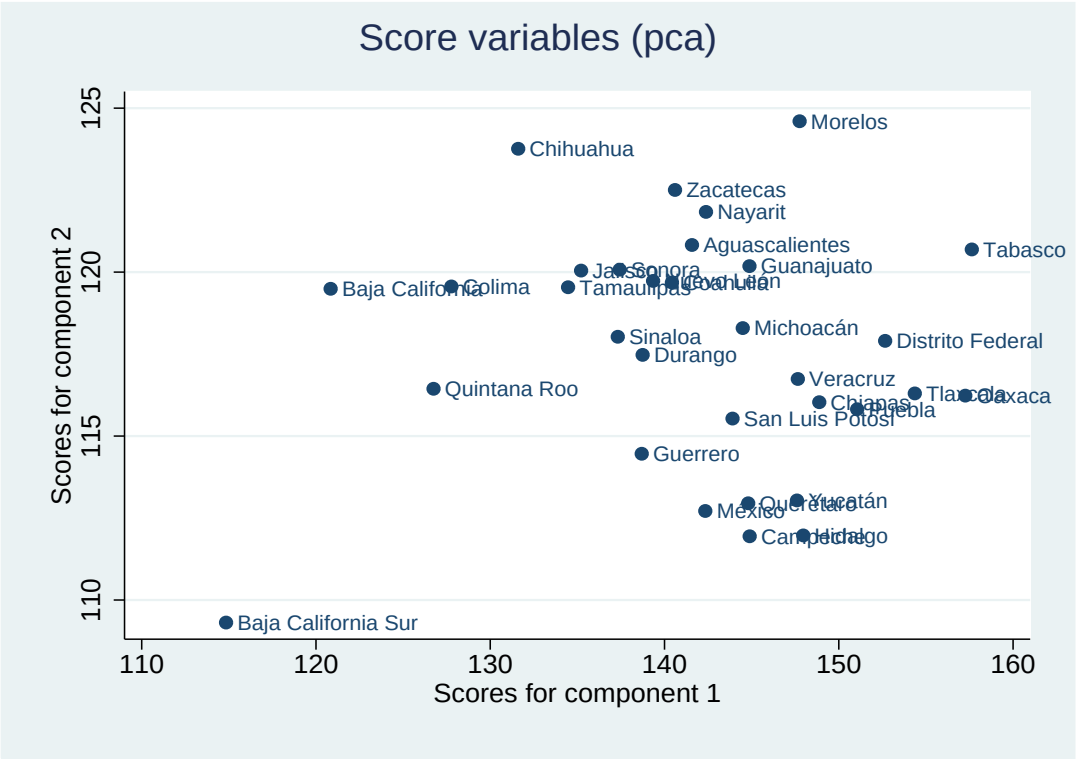
Región 1	Aguascalientes, Coahuila, Durango, Guanajuato, Guerrero, Michoacán, Morelos, Nayarit, San Luis Potosí, Zacatecas.
Región 2	Campeche, Chiapas, Distrito Federal, Hidalgo, México, Oaxaca, Puebla, Querétaro, Tabasco, Tlaxcala, Veracruz, Yucatán.
Región 3	Chihuahua, Jalisco, Nuevo León, Sinaloa, Sonora, Tamaulipas.
Región 4	Baja California, Quintana Roo, Colima.
Región 5	Baja California Sur.



Se puede observar que por cada región reportada (tanto para resultados con la matriz de varianzas como con la matriz de correlaciones), se encuentra un grupo de estados que se han mantenido en éstos y los análisis previos (Conglomerados). En la región uno encontramos a Aguascalientes, Nayarit, Durango, Michoacán y Guerrero (y para el caso con cinco regiones, se añaden Sinaloa, Sonora y Jalisco). Los estados de Oaxaca Puebla, Tabasco, Tlaxcala y Distrito Federal se mantuvieron en todo momento en una misma región. Lo mismo ocurrió con Campeche, Hidalgo, México, Querétaro, San Luis Potosí, Yucatán y Veracruz.

Todo lo anterior sugiere revisar el comportamiento de los estados con base en el análisis de componentes principales. Se espera que las regiones resalten en las gráficas de los dos primeros componentes que se presentan a continuación.

4.5.3. Gráfica de los dos primeros componentes principales obtenidos con la matriz de varianzas y covarianzas.



El porcentaje de varianza explicado por las primeras dos componentes es 88 %.

De la gráfica:

1. Se verifica el "*alejamiento*" de Baja California Sur, comportamiento que se preservó en todos los resultados.

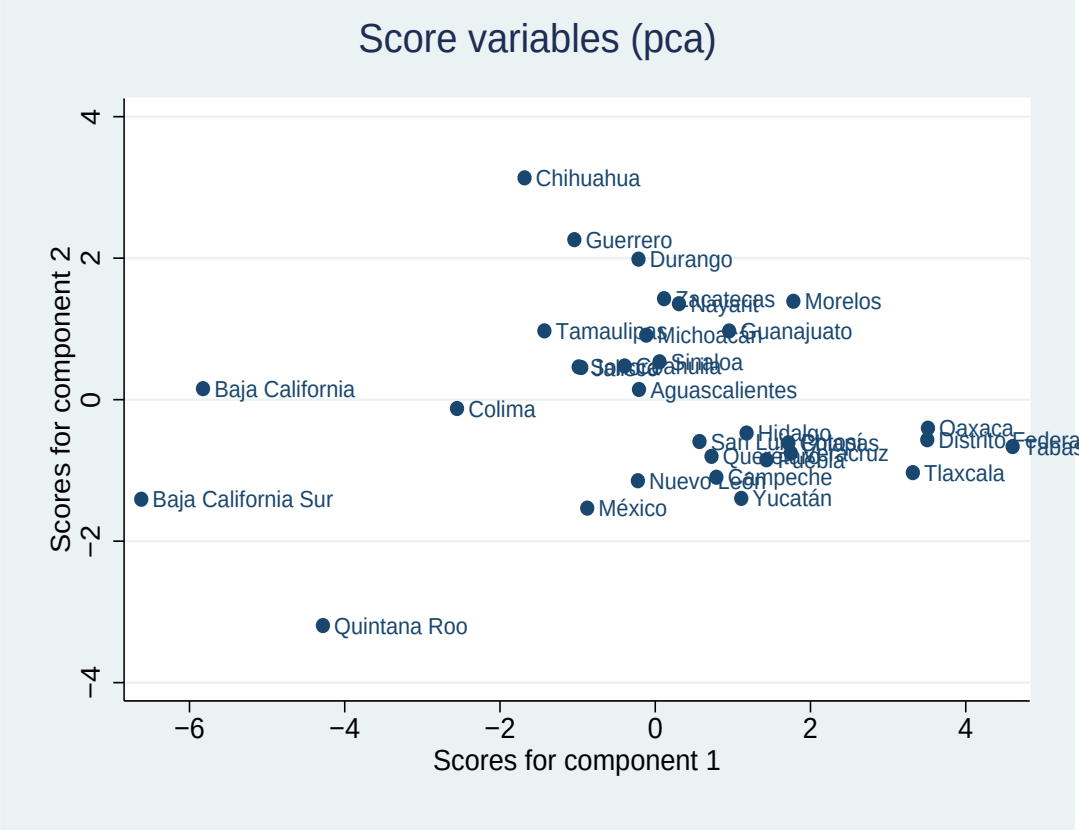
2. Es claro ver la cercanía de Baja California, Colima, Chihuahua y Quintana Roo, con lo cual, se verifica también el hecho que se presentó en los resultados al pertenecer éstos a una misma región.

3. A la derecha, se observan adyacentes los estados de Tabasco, Tlaxcala, Oaxaca y Distrito Federal, los cuales, se mantuvieron en una misma región.

4. En las zonas centro e inferior, se observa un grupo de estados que se asociaron en dos o tres regiones, según sea el caso en análisis.

5. En la zona central, también se tiene (no de manera tan clara) un grupo de estados que formaron parte de una misma región, a decir: Durango, Tamaulipas, Jalisco, Sonora, Sinaloa y Nuevo León.

4.5.4. Gráfica de los dos primeros componentes principales obtenidos con la matriz de correlaciones.



El porcentaje de varianza explicado por las primeras dos componentes es 64 %.

Nótese que:

1. Se corroboran los "*alejamientos*" de Baja California Sur (que se aisló para formar una región en el Análisis de Conglomerados y la mayoría de k - medias) y de Baja California, Colima y Quintana Roo.

2. De manera más clara (respecto a la gráfica anterior), se observan al centro los estados de: Durango, Tamaulipas, Jalisco, Sonora, Sinaloa y Nuevo León, los cuales, se mantuvieron juntos en una misma región en la gran parte de los análisis.

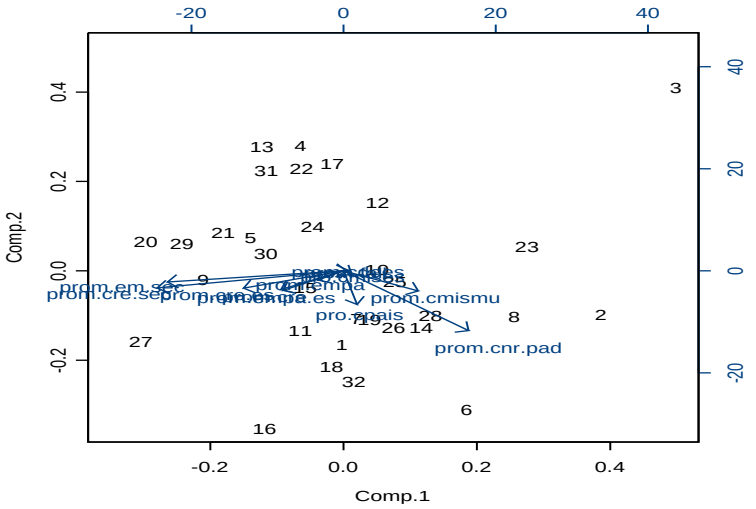
3. En la zona derecha de la gráfica se observa nuevamente un grupo de estados que estuvieron en la mayoría de los análisis, formando una región, a decir: Distrito Federal, Oaxaca, Tlaxcala y Tabasco[27].

4. En la parte central de la gráfica se observa otro grupo de estados: Campeche, Querétaro, Yucatán, Hidalgo, San Luis Potosí, Chiapas, Veracruz y México, los cuales formaron una región en prácticamente todos los análisis realizados (incluidos los análsis con las VNM2005 y VNM2006 sin ponderar).

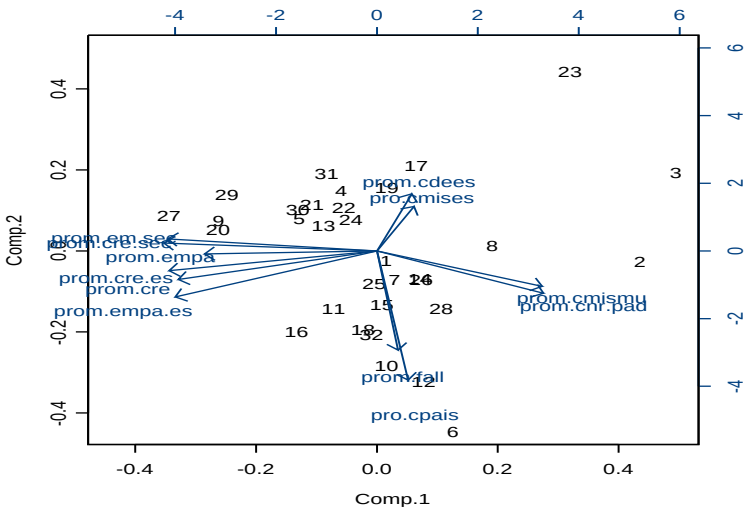
[27] Esta situación también se presentó en los análisis previos (con 2p variables) y para las VNM2005 y VNM2006 (sin ponderar).

4.5.5. Gráficas Biplot de los componentes principales obtenidos con el promedio aritmético de los indicadores

4.5.5.1. Biplot con la matriz de varianzas y covarianzas.



4.5.5.2. Biplot con la matriz de correlaciones.



Notación	**Variables (representadas con los vectores.)**
prom.empa	Promedio de empadronados.
prom.empa.es	Promedio de empadronados en el estado.
prom.em.sec	Promedio de empadronados en la sección.
prom.cre	Promedio de credencializados.
prom.cre.es	Promedio de credencializados en el estado.
prom.cre.sec	Promedio de credencializados en la sección.
prom.fall	Promedio de fallecidos.
prom.cnr.pad	Promedio de Cambios de domicilio no reportados en el padrón.
prom.cmismu	Promedio de cambios de domicilio al mismo municipio.
prom.cmises	Promedio de cambios de domicilio a otro municipio dentro del mismo estado.
prom.cdees	Promedio de cambios de domicilio a otro estado.
prom.cpais	Promedio de Cambios de domicilio a otro país.

Esencialmente, las variables conservan su comportamiento respecto a los Biplots de secciones previas, aunque no es clara su relación con los estados. Particularmente en el Biplot con la matriz de correlaciones, se observan indicadores correlacionados, específicamente con las parejas de indicadores: Cambios de país y Fallecidos; Cambios de domicilio en el mismo municipio y Cambios no reportados al Padrón; y Cambios de estado y Cambios dentro del mismo estado. Además, se observan entidades *alejadas* y que en la mayoría de los análisis conformaron una región, a decir: Baja California (2), Baja California Sur (3), Quintana Roo (23) y Colima (6); y aunque ello no indica que necesariamente pertenezcan a una misma región, sí presentan comportamientos distintos al resto de las entidades, y en ese sentido se puede interpretar que pertenecen a una misma región.

El comportamiento que presentaron las 32 entidades en este análisis mantuvo una gran semejanza y similaridad con los resultados obtenidos en las Verificaciones de 2005 y 2006 (sin ponderar), lo que es síntoma de una consistencia interna e inherente a las características electorales de los estados, en particular, con los siguientes grupos de estados (no está el total de las entidades federativas):

1.- Distrito Federal, Oaxaca, Tlaxcala, Puebla, Tabasco.
2.- Durango, Guerrero, Jalisco, Tamaulipas, Sonora, Sinaloa, Nuevo León.
3.- Baja California Sur, Baja California, Quintana Roo, Colima, Chihuahua.
4.- Campeche, Querétaro, San Luis Potosí, Hidalgo, Yucatán, México,
Chiapas, Veracruz.

Esta "*conducta*" refleja por un lado, que el promedio aritmético de los indicadores está cercano a ambos o lo que es igual, que ambos indicadores son similares, salvo por alguna cantidad no considerable. Por otro lado, puede interpretarse como el hecho de aplicarle un "*peso*" de un medio a cada indicador lo cual se traduce en darle la misma importancia a ambos. Finalmente, surge la pregunta: ¿Es correcto darle dicho peso a ambos indicadores?, o equivalentemente, ¿Cuál de ambas Verificaciones tendrá un mayor peso (o aporte) para la conformación de regiones? Al parecer, los resultados obtenidos hasta el momento sugieren que la VNM2006, sin embargo estas preguntas se intentarán responder en la sección siguiente, donde se tratarán nuevamente indicadores ponderados, pero ahora el *peso o factor de ponderación* no necesariamente será $\frac{1}{2}$ [28].

[28] La metodología para encontrar dichos ponderadores se detalla, precisamente, en la siguiente sección.

4.6. Resultados utilizando ponderaciones (1/4, 3/4).

En el Capítulo I se abordaron con detalle cada una de las Verificaciones que han sido realizadas. En particular, recordemos los aspectos del Diseño muestral aplicado en las Verificaciones de 2005 y 2006. En el caso de la VNM2005, el esquema de muestreo fue polietápico y estratificado. La determinación del tamaño de muestra y su distribución estuvo en función de: a) el nivel de inferencia requerido y b) el presupuesto disponible.

Dado el interés por alcanzar una inferencia a nivel estatal, se estableció un tamaño de muestra de 2,500 secciones. Debido a las restricciones presupuestales para poder llevar a cabo una verificación de esta magnitud, la opción elegida fue emplear la misma muestra de secciones que se habían visitado en la Verificación Nacional Muestral de 2003[29]. Al emplear la misma muestra de secciones que en el 2003, se aprovechó la información del recorrido cartográfico realizado dos años antes, con esto se tuvo un ahorro en la Encuesta de Cobertura, específicamente en la elaboración del marco muestral para la segunda etapa de muestreo (relación de manzanas con viviendas habitadas).

Con esta muestra de secciones y una confianza de 95 %, a nivel nacional se estimó que las precisiones estadísticas serían: 0.3 % para Empadronados, 0.8 % para Credencializados en la sección y 0.5 % para Residentes en la sección del padrón. Mientras que a nivel estatal se estimó que las precisiones estadísticas estarían alrededor de 2.0 % para Empadronados, 5.5 % para Credencializados en la sección y 3.0 % para Residentes en la sección.

[29] IFE-RFE, "Verificación Nacional Muestral 2005. Objetivos, Indicadores y Etapas de selección de la muestra. Segunda versión", 6 de abril de 2005.

Para la VNM2006, el esquema de muestreo fue polietápico y estratificado. El tamaño de muestra y su distribución quedó determinado por el nivel de inferencia requerido[30]. En ese sentido, se determinó el número de secciones a seleccionar por estado y tipo de distrito, para lo cual se empleó (en ambos niveles de inferencia) el efecto del diseño (*Deff*) de la VNM2003.

El tamaño de muestra se fijó en 2,940 secciones y a partir de éste, de la distribución establecida y de una confianza de 95 %, se estimó que, a nivel nacional, las precisiones estadísticas serían: 0.3 % para Empadronados, 0.8 % para Credencializados en la sección y 0.5 % para Residentes en la sección. Y para las seis categorías de distritos las precisiones estuvieron alrededor de: 1.0 %, 2.0 % y 1.5 % respectivamente.

Como puede verse, fue en la VNM2006 donde se empleó un mayor tamaño de muestra y se lograron obtener estimaciones con una mejor precisión. Por ello, en esta sección se consideró la siguiente combinación lineal de los indicadores:

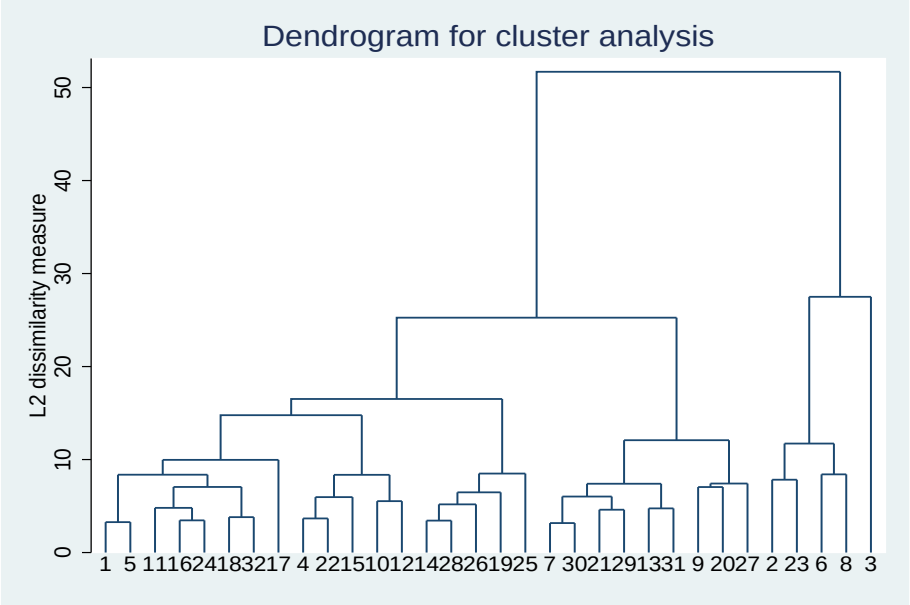
$$\omega^{i,j} = \frac{1}{4}\omega^{i,j}_{2005} + \frac{3}{4}\omega^{i,j}_{2006}$$

Además, si a las precisiones se le añade el hecho de que en los resultados de secciones anteriores, los datos obtenidos en VNM2006 impactaron en mayor proporción al momento de formar las regiones, entonces tiene sentido el proporcionar un mayor *peso* a la VNM2006, de tal forma que si hubiese un cambio en el comportamiento de las regiones, éste sería más adecuado. Los resultados se presentarán en la siguiente secuencia: Análisis de Conglomerados, gráficas de Componentes Principales (con la matriz de varianzas y covarianzas y la matriz de correlaciones) y Biplots.

[30] Nacional, estatal y categorías de distritos.

4.6.1. Análisis de Conglomerados.

4.6.1.1. Análisis de Conglomerados con *liga completa*, tomando como medida de disimilaridad la norma euclideana.

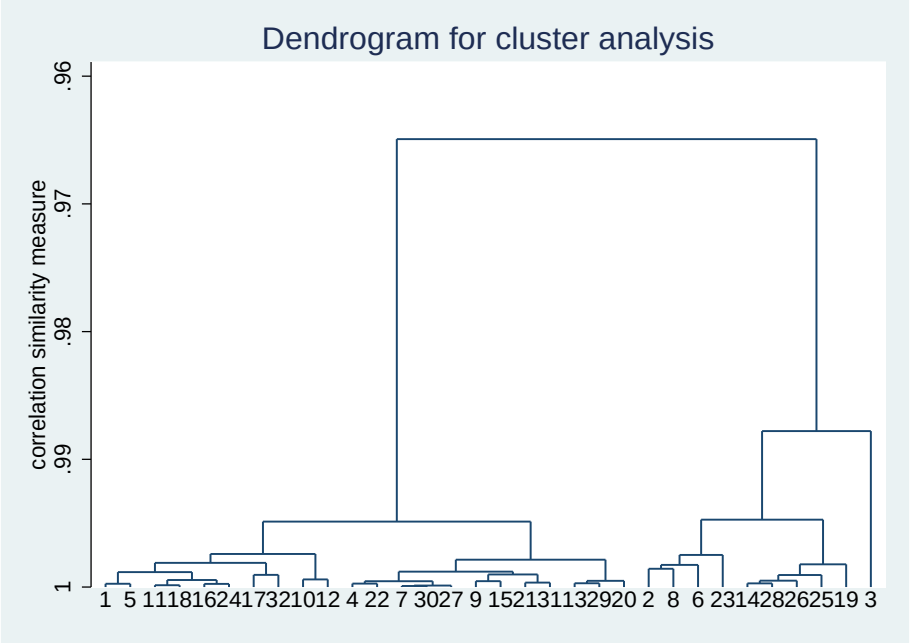


Realizando el corte correspondiente se obtienen las siguientes regiones:

Región 1	(Aguascalientes, Coahuila, Guanajuato, Michoacán, San Luis Potosí, Nayarit, Zacatecas, Morelos), (Campeche, Querétaro, México, Durango, Guerrero).
Región 2	Jalisco, Tamaulipas, Sonora, Nuevo León, Sinaloa.
Región 3	Chiapas, Veracruz, Puebla, Tlaxcala, Hidalgo, Yucatán, Distrito Federal, Oaxaca, Tabasco.
Región 4	Baja California, Quintana Roo, Colima, Chihuahua.
Región 5	Baja California Sur.

Nótese que todas las regiones (en particular la región 1, 3, 4 y 5) mantienen su estructura respecto al análisis de la sección previa y más aún, con los resultados obtenidos para la VNM2006. Pero nótese con base en el dendrograma, que la región 1 podría separarse en dos, como lo muestran los paréntesis.

4.6.1.2. Análisis de Conglomerados con *liga completa*, tomando como medida de similaridad la matriz de correlaciones.

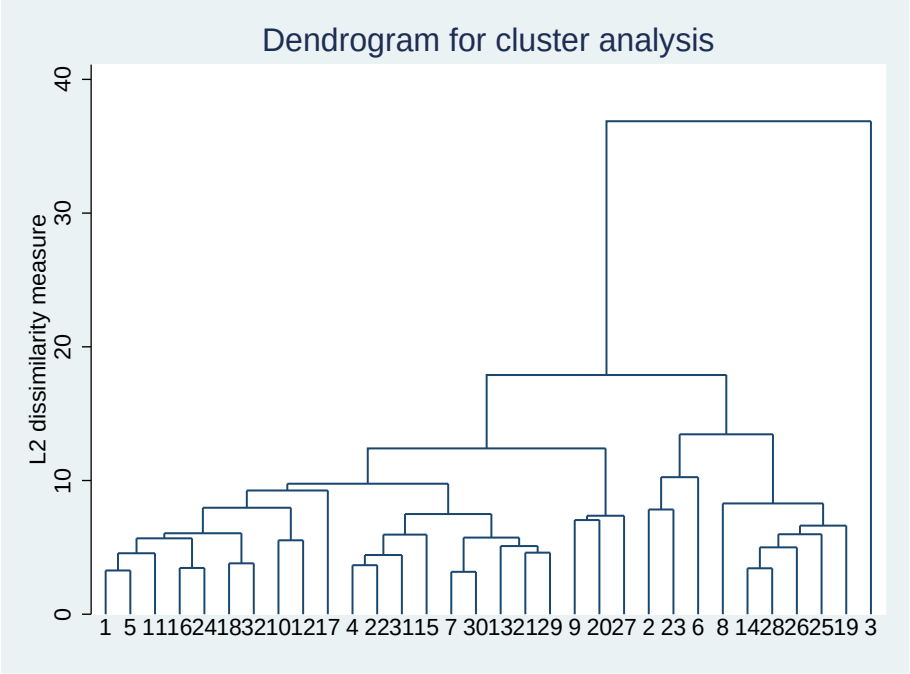


Al realizar el corte correspondiente respecto a la medida de similaridad, se obtienen las siguientes regiones:

Región 1	Aguascalientes, Coahuila, Guanajuato, Nayarit, Michoacán, San Luis Potosí, Morelos, Zacatecas, Durango, Guerrero.
Región 2	Campeche, Querétaro, Chiapas, Veracruz, Tabasco, Distrito Federal, México, Puebla, Yucatán, Hidalgo, Tlaxcala, Oaxaca.
Región 3	Baja California, Chihuahua, Colima, Quintana Roo.
Región 4	Jalisco, Tamaulipas, Sonora, Sinaloa, Nuevo León.
Región 5	Baja California Sur.

El comportamiento de Baja California Sur prevalece. La estructura de las regiones 1, 3 y 4 es comparable con los resultados de 4.5.1.1 así como las regiones 2, 4 y 5 con el correspondiente análisis en la VNM2006. Nótese que se ha mantenido el grupo de estados del norte: Sonora, Sinaloa, Nuevo Léon, Jalisco y Tamaulipas.

4.6.1.3. Análisis de Conglomerados con *liga promedio*, tomando como medida de disimilaridad la norma euclideana.

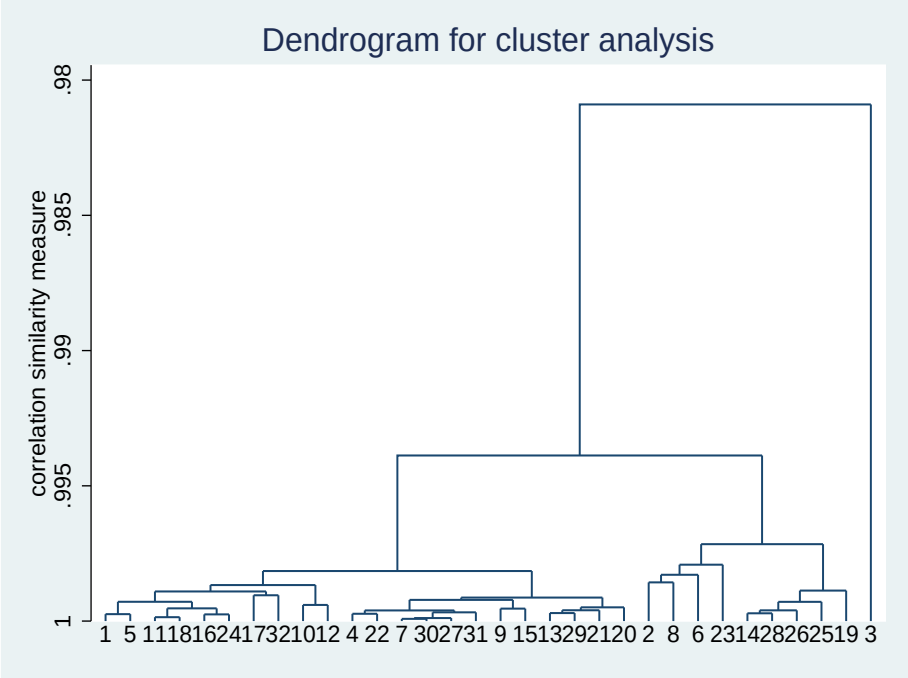


Al ejecutar el corte correspondiente respecto a la medida de disimilaridad, resultan las siguientes regiones:

Región 1	(Aguascalientes, Coahuila, Guanajuato, Michoacán, San Luis Potosí, Nayarit, Zacatecas, Durango, Guerrero, Morelos.
Región 2	Campeche, Querétaro, Yucatán, México, Chiapas, Veracruz, Hidalgo, Puebla, Tlaxcala).
Región 3	Distrito Federal, Oaxaca, Tabasco.
Región 4	Baja California, Quintana Roo, Colima.
Región 5	Chihuahua, Jalisco, Tamaulipas, Sonora, Sinaloa, Nuevo León.
Región 6	Baja California Sur.

Las regiones en esencia se conservan y se tiene un efecto de *encadenamiento*. Nótese de nuevo la cohesión de los estados del norte (región 5) y la que se presentan en las regiones 2, 3 y 4. Incluso, las regiones 1 y 2 pueden verse como una sola. En general, ya se tiene una *estructura* en las regiones con dirección a que la VNM2006 tiene mayor impacto.

4.6.1.4. Análisis de Conglomerados con *liga promedio*, tomando como medida de similaridad la matriz de correlaciones.

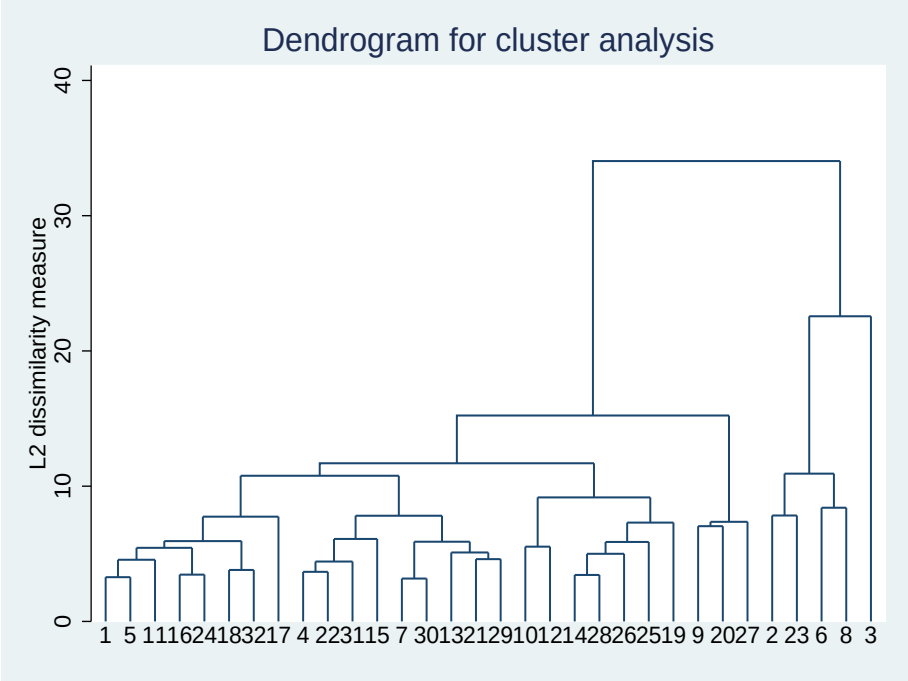


Con el corte correspondiente, se obtienen las siguientes regiones:

Región 1	Aguascalientes, Coahuila, Guanajuato, Nayarit, Michoacán, San Luis Potosí, Morelos, Zacatecas, Durango, Guerrero.
Región 2	Campeche, Querétaro, Chiapas, Veracruz, Tabasco, Yucatán Distrito Federal, México, Hidalgo, Tlaxcala, Puebla, Oaxaca.
Región 3	Baja California, Chihuahua, Colima, Quintana Roo.
Región 4	Jalisco, Tamaulipas, Sonora, Sinaloa, Nuevo León.
Región 5	Baja California Sur.

A excepción de la región 2, las restantes comparten esencialmente la misma estructura con las regiones de los análisis previos. Sin embargo, y debido a la cercanía reflejada en el dendrograma, se puede considerar que las regiones 1 y 2 para este caso, forman una misma.

4.6.1.5. Análisis de Conglomerados con *liga peso - promedio ponderado* (Media de grupos), tomando como medida de disimilaridad la norma euclideana.

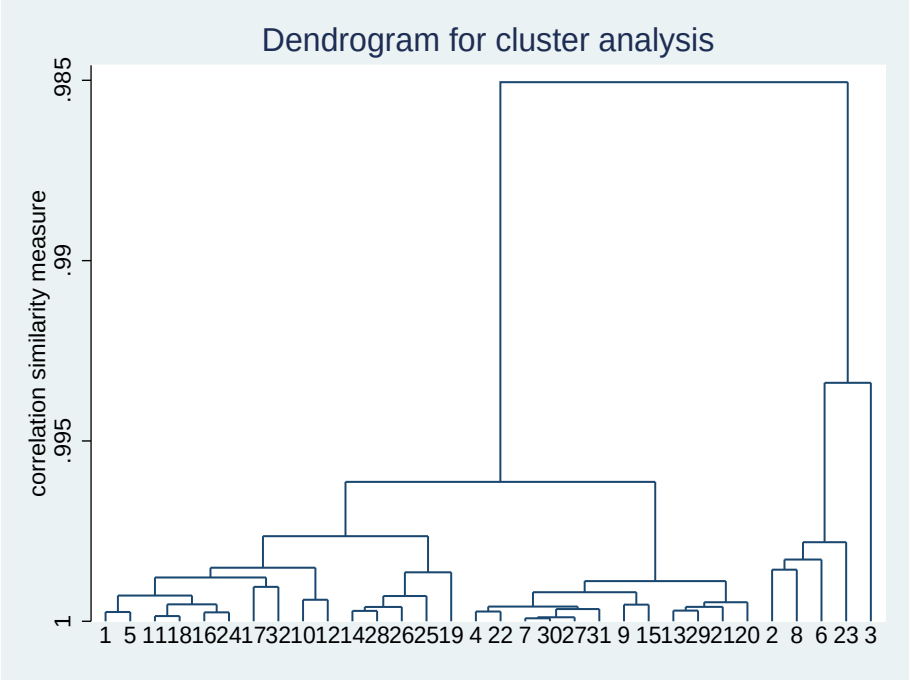


Al realizar el corte correspondiente respecto a la medida de disimilaridad, se obtienen las siguientes regiones:

Región 1	(Aguascalientes, Coahuila, Guanajuato, Michoacán, San Luis Potosí, Nayarit, Zacatecas, Morelos), (Campeche, Querétaro, Yucatán, México, Chiapas, Puebla, Tlaxcala).
Región 2	Durango, Guerrero, Jalisco, Tamaulipas, Sonora, Sinaloa, Nuevo León.
Región 3	Distrito Federal, Oaxaca, Tabasco.
Región 4	Baja California, Quintana Roo, Colima, Chihuahua.
Región 5	Baja California Sur.

Los resultados son similares a los obtenidos en la subsección 4.6.1.1. salvo por la región tres. Debido a la cercanía y en función de la altura del corte, puede considerarse que la región uno se separa en dos regiones.

4.6.1.6. Análisis de Conglomerados con *liga peso - promedio ponderado* (Media de grupos), tomando como medida de similaridad la matriz de correlaciones.

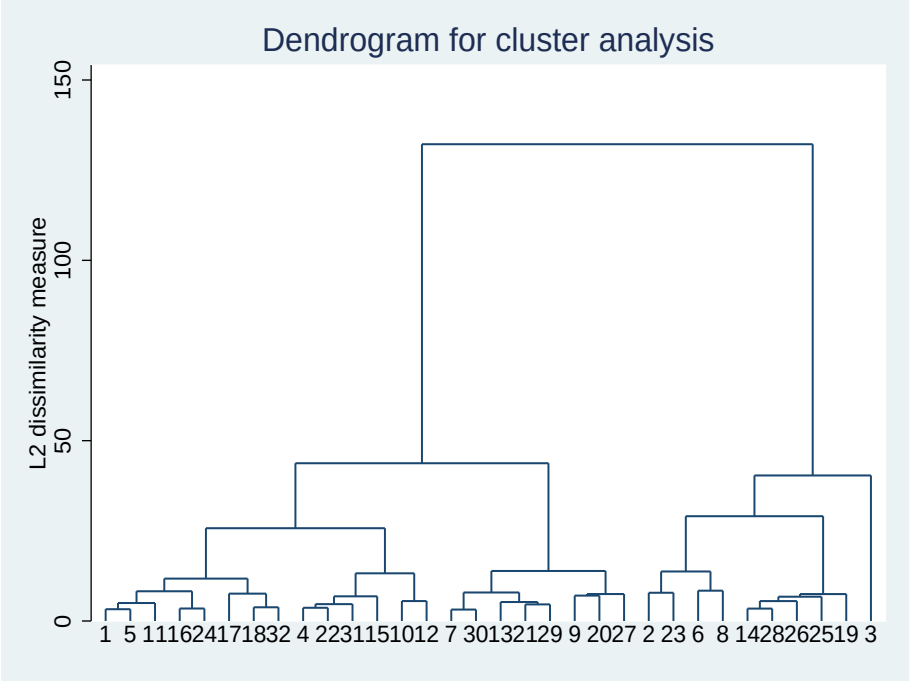


Al ejecutar el corte correspondiente se obtienen las siguientes regiones:

Región 1	(Aguascalientes, Coahuila, Guanajuato, Nayarit, Michoacán, San Luis Potosí, Morelos, Zacatecas, Durango, Guerrero.
Región 2	Jalisco, Tamaulipas, Sonora, Sinaloa, Nuevo León).
Región 3	Campeche, Querétaro, Chiapas, Veracruz, Tabasco, Yucatán, Distrito Federal, México, Hidalgo, Tlaxcala, Puebla, Oaxaca.
Región 4	Baja California, Chihuahua, Colima, Quintana Roo.
Región 5	Baja California Sur.

Esencialmente, se obtienen las mismas regiones que en 4.5.1.4. Además, el comportamiento de Baja California y Quintana Roo se mantiene al permanecer en una misma región; y lo mismo sucede con Baja California Sur. Además, pueden distinguirse incluso cuatro regiones, pues con base en el dendrograma, las regiones 1 y 2 podrían unirse.

4.6.1.7. Análisis de Conglomerados con el *Método de Ward*, tomando como medida de disimilaridad la norma euclideana.

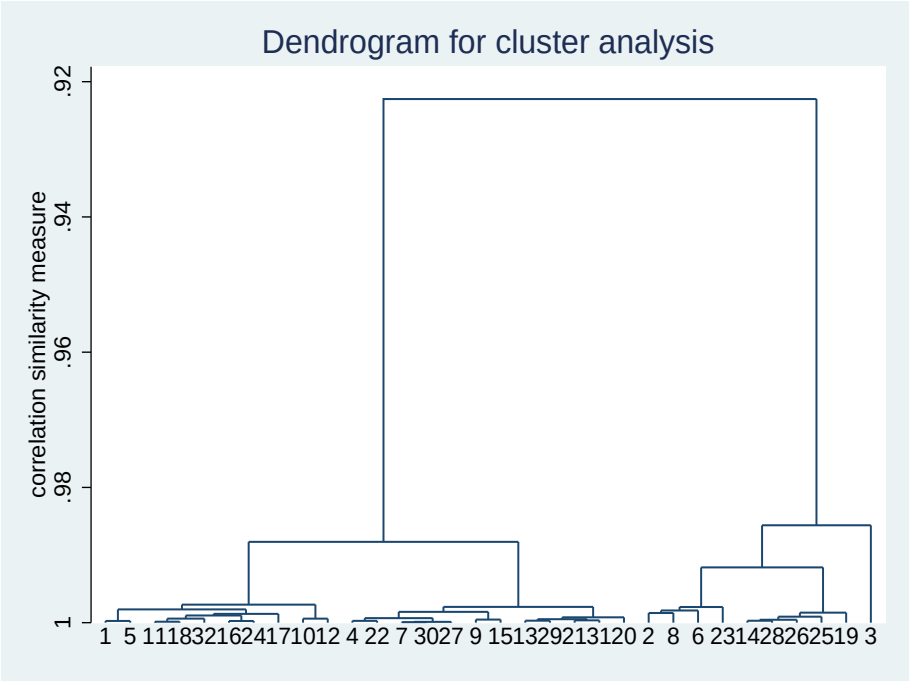


Realizando el corte correspondiente se obtienen las siguientes regiones:

Región 1	(Aguascalientes, Coahuila, Guanajuato, Michoacán, San Luis Potosí, Morelos, Nayarit, Zacatecas), (Campeche, Querétaro, Yucatán, México, Durango, Guerrero).
Región 2	Chiapas, Yucatán, Hidalgo, Puebla, Tlaxcala, Distrito Federal, Oaxaca, Tabasco.
Región 3	Baja California, Quintana Roo, Colima, Chihuahua.
Región 4	Jalisco, Tamaulipas, Sonora, Sinaloa, Nuevo León.
Región 5	Baja California Sur.

Es importante mencionar que las regiones han mantenido una *estuctura interna* en cuanto a los estados que las conforman. Por ejemplo, la cohesión del grupo de estados del norte formado por: Tamaulipas, Jalisco, Sonora, Sinaloa y Nuevo León. Lo mismo sucede con los estados de Puebla, Tlaxcala, Tabasco y Distrito Federal.

4.6.1.8. Análisis de Conglomerados con el *Método de Ward*, tomando como medida de similaridad la matriz de correlaciones.

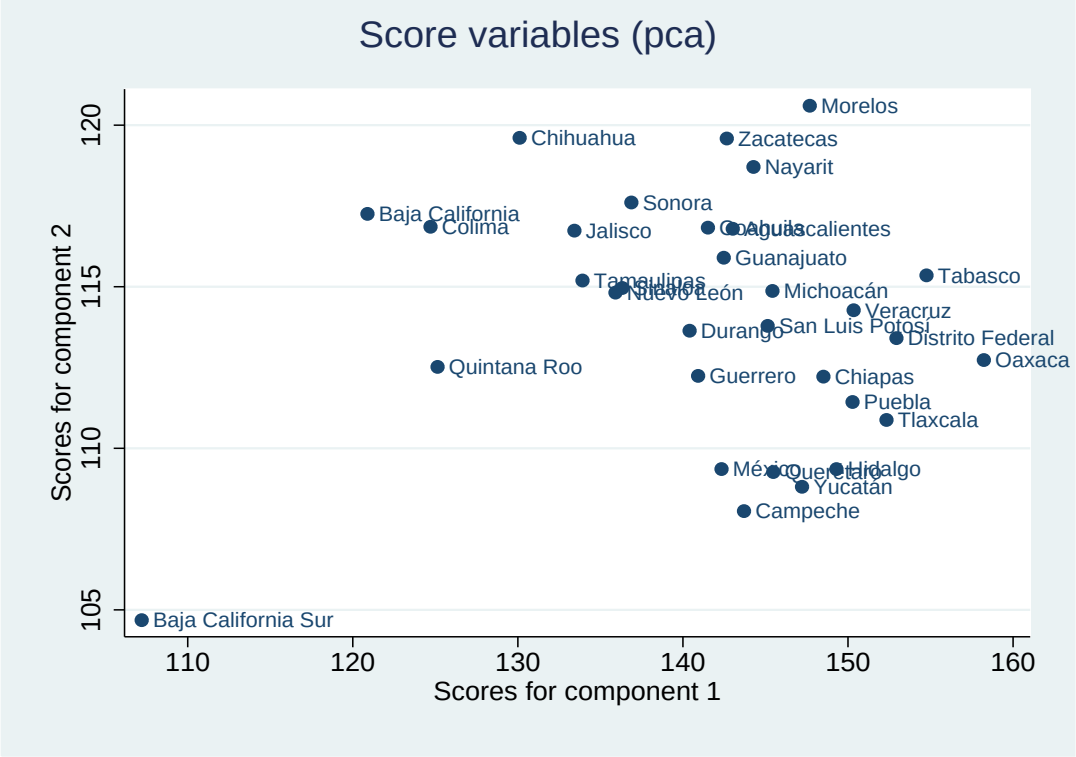


Al realizar el corte correspondiente se obtienen las siguientes regiones:

Región 1	Aguascalientes, Coahuila, Guanajuato, Nayarit, Zacatecas, Michoacán, San Luis Potosí, Morelos, Durango, Guerrero.
Región 2	Campeche, Querétaro, Chiapas, Veracruz, Tabasco, Distrito Federal, México, Hidalgo, Tlaxcala, Puebla, Yucatán, Oaxaca.
Región 3	Baja California, Chihuahua, Colima, Quintana Roo.
Región 4	Jalisco, Tamaulipas, Sonora, Sinaloa, Nuevo León.
Región 5	Baja California Sur.

Nótese que si comparamos los resultados obtenidos en esta sección con los correspondientes obtenidos con la VNM2006 (sin ponderar), se observan resultados sumamente similares en las regiones propuestas. Esto puede verse con la región que contiene a estados del norte de la República Mexicana (región 2), asimismo se verifica que Baja California Sur se comporta de manera diferente al resto de las entidades.

4.6.2. Gráfica de los dos primeros componentes principales obtenidos con la matriz de varianzas y covarianzas.



El porcentaje de varianza explicado por las primeras dos componentes es 90 %.

De la gráfica:

1. Se presenta nuevamente el comportamiento de Baja California Sur. Esto, como se ha visto, también ocurre con las VNM2006 y con el promedio aritmético de los indicadores (Sección 4.5).

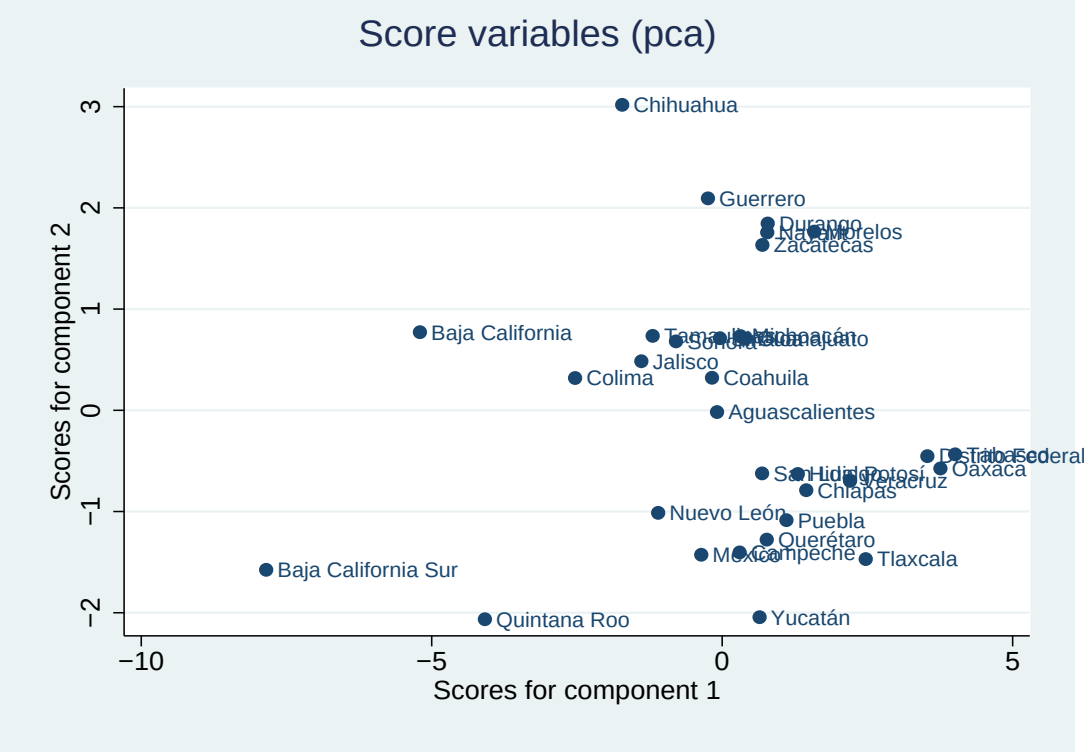
2. Es claro ver la cercanía de Baja California, Colima, Chihuahua y Quintana Roo[31], con lo cual se apoyan aún más los resultados obtenidos previamente (y en el análisis de la VNM2006 y con el promedio de los indicadores) al pertenecer éstos a una misma región.

[31] Principalmente respecto a la primer componente.

3. En la zona inferior - derecha se observa la cohesión de los estados: México, Hidalgo, Yucatán, Puebla, Tlaxcala y Oaxaca entre otros, quienes formaron una región en los resultados, no solamente en esta sección, también con los análisis de la VNM2006 (Sección 4.2), con todos los indicadores (Sección 4.4) y con el promedio de ellos (Sección 4.5).

4. En la zona central, se tiene el grupo de estados del norte de la República, los cuales conformaron una región en todos los análisis: Jalisco, Tamaulipas, Sonora, Sinaloa y Nuevo León[32].

4.6.3. Gráfica de los dos primeros componentes principales obtenidos con la matriz de correlaciones.



El porcentaje de varianza explicado por las primeras dos componentes es 64 %.

[32]Comportamiento que también se detectó en los análisis de la VNM2006 (Sección 4.2), con todos los indicadores (Sección 4.4) y con el promedio de ellos (Sección 4.5).

Nótese que:

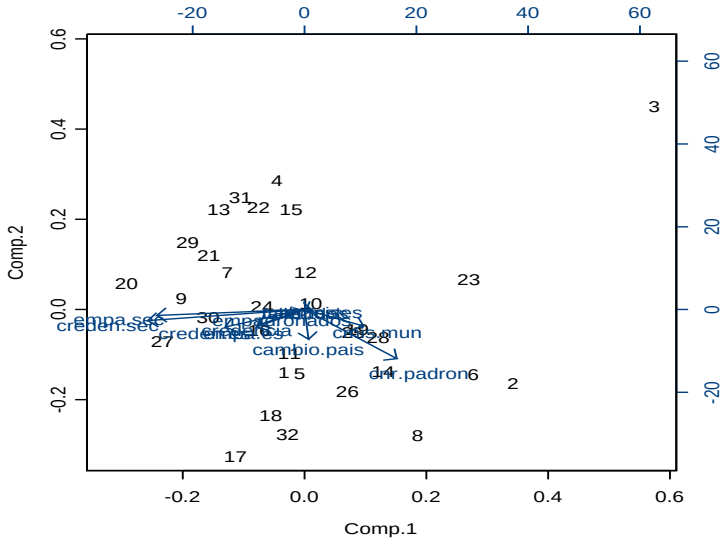
1. Se verifica el "*alejamiento*" de Baja California Sur, comportamiento que se presentó en todos los resultados.

2. En el centro de la gráfica se observan los estados de: Tamaulipas, Jalisco, Sonora, Sinaloa y Nuevo León, los cuales, se mantuvieron juntos en una misma región en gran parte de los análisis[33].

3. En la zona derecha de la gráfica se observa nuevamente un grupo de estados que estuvieron en la mayoría de los análisis formando una región: Distrito Federal, Oaxaca, Puebla, Yucatán, Tlaxcala y Tabasco.

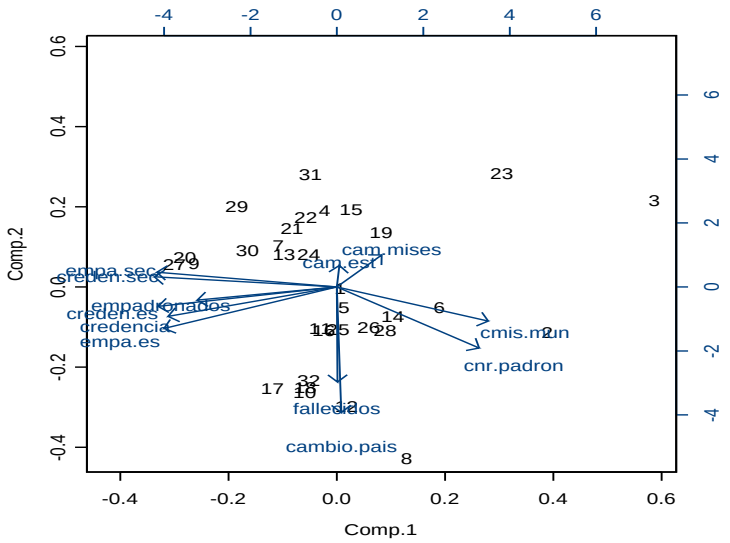
4.6.4. Gráficas Biplot para los componentes principales obtenidos con el promedio aritmético de los indicadores

4.6.4.1. Biplot con la matriz de varianzas y covarianzas.



[33]Esta situación también se presentó en los análisis con todos los indicadores (2p variables), el promedio de indicadores, y para las VNM2005 y VNM2006 (sin ponderar).

4.6.4.2. Biplot con la matriz de correlaciones.



Notación	Variables (representadas con los vectores.)
empadronados	Empadronados.
empa.es	Empadronados en el estado.
empa.sec	Empadronados en la sección.
credencia	Credencializados.
creden.es	Credencializados en el estado.
creden.sec	Credencializados en la sección.
fallecidos	Fallecidos.
cnr.pad	Cambios de domicilio no reportados en el padrón.
cmismu	Cambios de domicilio al mismo municipio.
cam.cmises	Cambios de domicilio a otro municipio dentro del mismo estado.
cam.es	Cambios de domicilio a otro estado.
cambio.pais	Cambios de domicilio a otro país.

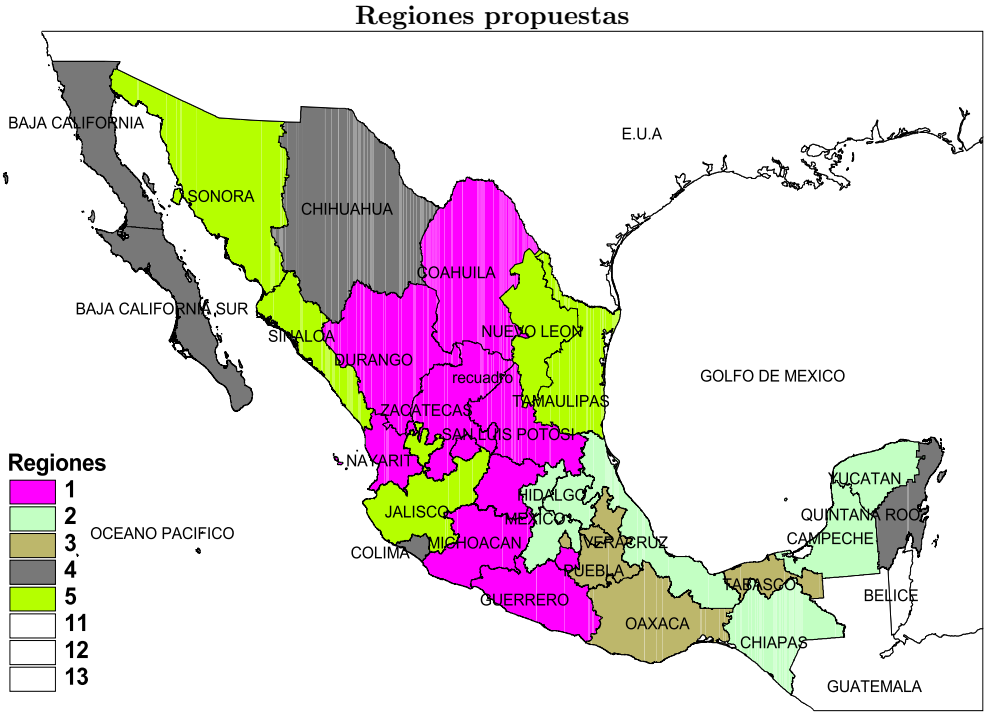
En el biplot con la matriz de correlaciones se observan tres grupos de variables claramente correlacionadas, entre ellos están: Cambios de domicilio en el mismo municipio y Cambios no reportados en el Padrón; y Fallecidos y Cambios de aís. Nótese que con este último se relacionan los estados de Durango (10), Morelos (17) y Zacatecas (32). Se verifica el comportamiento de los estados de Baja California Sur (3), Baja California (2), Quintana Roo (23) y Chihuahua (8).

4.7. Una propuesta de regionalización.

Con todos los resultados que se han analizado, es claro que la Verificación Nacional Muestral de 2006 tiene un mejor *diseño de muestra* y por ende un mayor *peso* en la conformación de las regiones. De esta manera, se cuenta con evidencia para afirmar (o al menos apoyar) que se observan *estructuras* fijas en el comportamiento de los estados. Se corrobora la cohesión de un grupo de estados del norte, de un grupo en el centro y a diferencia de la regionalización propuesta por CONAPO con base en el IM, no se presenta la cohesión (quizás esperada) en los estados de la Península de Yucatán.

Con base en los resultados y al hecho de que estamos bajo el supuesto de términos electorales, se proponen las siguientes regiones:

Región 1	Aguascalientes, Coahuila, Guanajuato, Nayarit, Zacatecas, Michoacán, San Luis Potosí, Morelos, Durango, Guerrero.
Región 2	Campeche, Chiapas, Querétaro, Veracruz, México, Yucatán, Hidalgo.
Región 3	Distrito Federal, Tlaxcala, Tabasco, Puebla, Oaxaca.
Región 4	Baja California, Chihuahua, Colima, Quintana Roo, Baja California Sur.
Región 5	Sonora, Sinaloa, Nuevo León, Jalisco, Tamaulipas.



Es claro que estas regiones no coinciden con la regionalización propuesta por CONAPO con base en el Índice de Marginación. Además, se puede observar que los resultados obtenidos con la VNM2006 impactan en mayor grado en la conformación de las regiones propuesta.

Con este análisis y comentarios no se pretende desacreditar la regionalización propuesta por CONAPO, sino por el contrario, recordemos que el objetivo de este proyecto es realizar una propuesta de regiones con base en indicadores electorales, y es en este sentido que se confirma que el comportamiento electoral en el República Mexicana es completamente diferente al comportamiento de las zonas marginadas. Por lo cual, se recomienda que para futuras Verificaciones o análisis que involucren aspectos electorales, considerar regiones conformadas con base a criterios electorales.

Capítulo 5

Análisis y validación de las regiones propuestas.

En este capítulo se exhibirán:

1. Los resultados de un Análisis Discriminante con cada modalidad abordada en el capítulo anterior (a excepción de los indicadores ponderados[1]):
   - La Verificación del 2005,
   - La Verificación del 2006,
   - Los indicadores de 2005 y 2006,
   - El promedio de los indicadores y
   - La ponderación $(1/4, 3/4)$ para los indicadores.
2. Los resultados de un Análisis Discriminante para los datos de la VNM2003, que fue la última en donde se realizaron inferencias a nivel regional.
3. Un resumen con las principales medidas descriptivas de las regiones propuestas así como de las entidades que las conforman.

[1] Lo cual se debe a que en los resultados con los **indicadores ponderados** no se observó estructura alguna en las regiones. Se optó por analizar las modalidades en donde se presentaron regiones similares a la propuesta.

5.1. Resultados del Análisis Discriminante.

Para realizar el Análisis Discriminante se utilizó el software S-PLUS. Dicho análisis proporciona información sobre la consistencia interna de las regiones propuestas, que se presentan a continuación:

Región 1	Aguascalientes, Coahuila, Guanajuato, Nayarit, Zacatecas, Michoacán, San Luis Potosí, Morelos, Durango, Guerrero.
Región 2	Campeche, Chiapas, Querétaro, Veracruz, México, Yucatán, Hidalgo.
Región 3	Distrito Federal, Tlaxcala, Tabasco, Puebla, Oaxaca.
Región 4	Baja California, Chihuahua, Colima, Quintana Roo, Baja California Sur.
Región 5	Sonora, Sinaloa, Nuevo León, Jalisco, Tamaulipas.

Para cada una de las modalidades mencionadas se exhibirán:

1. Prueba de homogeneidad en la estructura de covarianzas,
2. Las pruebas T^2 de Hotelling para diferencias entre vectores de medias,
3. Las distancias de Mahalanobis entre regiones, y
4. La tabla de clasificación *plug-in* para discriminación de los estados.

Para presentar estos resultados, se procedió a *generar* una nueva variable categórica en cada base de datos (VNM2005, VN2006, etc.). Ésta indica la *región final* propuesta en que se encuentra cada entidad federativa y que puede verse en la tabla previa. Asimismo, fue la variable que jugó el papel de variable dependiente en el Análisis Discriminante. A reserva de que se indique lo contrario, en todas las pruebas se considera el nivel de significancia $\alpha = 0.05$.

5.1.1. Resultados con la VNM2005.

A continuación se presentan los resultados del Análisis Discriminante con la VNM2005.

5.1.1.1. Prueba de homogeneidad en la estructura de covarianzas.

	Estadístico	Df	P-value
Box. M	253.822	312	0.99
adj. M	-6.5158	312	1.00

Es claro que la prueba no es significativa, es decir, la estructura de covarianzas no es la misma. Sin embargo, en este proyecto se trabajó bajo el supuesto de que la estructura de covarianzas es la misma pues de lo contrario, los parámetros a estimar serían demasiados.

5.1.1.2. Pruebas T^2 de Hotelling para diferencias entre medias.

	Estadístico F	Df1	Df2	P-value
G1-G2	3.3924	12	16	0.01237
G1-G3	4.9611	12	16	0.00185
G1-G4	3.7824	12	16	0.00742
G1-G5	1.6942	12	16	0.16073
G2-G3	2.5071	12	16	0.04407
G2-G4	7.9186	12	16	0.00012
G2-G5	3.2659	12	16	0.01470
G3-G4	10.313	12	16	0.00002
G3-G5	5.9168	12	16	0.00069
G4-G5	2.2977	12	16	0.06087

Para estas pruebas tenemos H_0*: Los dos grupos con todos los elementos en la escala tienen el mismo vector de medias.* Nótese que todas las pruebas son significativas a excepción de la prueba para los Grupos uno y cinco. En la siguiente

tabla se corrobora que esta diferencia no significativa corresponde a la mínima distancia de Mahalanobis.

5.1.1.3. Distancias de Mahalanobis entre regiones.

	G1	G2	G3	G4	G5
G1	0.00	16.6836	30.1388	22.9786	10.2928
G2		0.00	17.4067	54.9778	22.6747
G3			0.00	83.5393	47.9268
G4				0.00	18.6117
G5					0.00

5.1.1.4. Tabla de clasificación *plug-in* para discriminación de estados[2]

	G1	G2	G3	G4	G5	Error
G1	8	0	0	0	2	0.200
G2	0	7	0	0	0	0.000
G3	0	0	5	0	0	0.000
G4	0	0	0	5	0	0.000
G5	1	0	0	0	4	0.200

Como puede verse, hay dos estados propuestos en la Región 1 que se clasifican en el Grupo cinco; y un estado propuesto en la Región 5 que se clasifica en el Grupo uno. Los dos primeros son Aguascalientes y Coahuila, y el tercero es Tamaulipas. El caso de Aguascalientes no concuerda con los resultados del capítulo anterior, pues en la mayoría de éstos, se clasifica junto con Guanajuato, Nayarit, Zacatecas, etc. Lo mismo ocurre con Tamaulipas. Sin embargo, la clasificación de prácticamente todos los estados (salvo por los tres mencionados) coincide con las regiones propuestas, lo cual indica que la estructura interna de la regionalización propuesta se mantiene en la VNM2005.

[2]La tabla de clasificación plug-in se forma con la predicción de los datos originales en los grupos, según sea la función discriminante.

5.1.2. Resultados con la VNM2006.

A continuación se exhiben los resultados del Análisis Discriminante con la VNM2006.

5.1.2.1. Prueba de homogeneidad en la estructura de covarianzas.

	Estadístico	Df	P-value
Box. M	239.3022	312	0.99
adj. M	-6.1431	312	1.00

Nuevamente la prueba no es significativa, es decir, la estructura de covarianzas estadísticamente no es la misma. Se recuerda que en este proyecto se trabajó bajo el supuesto de homogeneidad en la estructura de covarianzas pues de lo contrario, los parámetros a estimar serían demasiados.

5.1.2.2. Pruebas T^2 de Hotelling para diferencias entre medias.

	Estadístico F	Df1	Df2	P-value
G1-G2	4.6169	12	16	0.0027
G1-G3	6.4074	12	16	0.0004
G1-G4	12.924	12	16	0.0000
G1-G5	2.1569	12	16	0.07599
G2-G3	0.9209	12	16	0.5491
G2-G4	22.954	12	16	0.0000
G2-G5	7.1924	12	16	0.0002
G3-G4	24.049	12	16	0.0000
G3-G5	8.902	12	16	0.00005
G4-G5	5.295	12	16	0.0013

Se contrasta H_0*: Los dos grupos con todos los elementos en la escala tienen el mismo vector de medias*. Nótese que todas las pruebas son significativas a excepción de la prueba para los parejas de Grupos uno y cinco; y Grupos dos con tres.

5.1.2.3. Distancias de Mahalanobis entre regiones.

	G1	G2	G3	G4	G5
G1	0.00	22.7056	38.9254	78.5157	13.1031
G2		0.00	6.3941	159.37	49.936
G3			0.00	194.7969	72.1066
G4				0.00	42.888
G5					0.00

5.1.2.4. Tabla de clasificación *plug-in* para discriminación de estados.

	G1	G2	G3	G4	G5	Error
G1	10	0	0	0	0	0.000
G2	0	6	1	0	0	0.1428
G3	0	1	4	0	0	0.2000
G4	0	0	0	5	0	0.0000
G5	0	0	0	0	5	0.0000

Hay un estado propuesto en la Región 2 y que se clasificó en el Grupo 3: Querétaro; y un estado propuesto en la Región 3, que se clasifica en el Grupo dos: Puebla. Pero nuevamente es de importancia saber que a excepción de los dos mencionados, la discriminación de los estados vía el Análisis Discriminante no presenta más errores para las regiones propuestas, lo cual indica que la estructura de regionalización se mantiene en la VNM2006.

5.1.3. Resultados con todos los indicadores.

A continuación se exhiben los resultados del Análisis Discriminante con todos los indicadores.

5.1.3.1. Prueba de homogeneidad en la estructura de covarianzas.

	Estadístico	Df	P-value
Box. M	1176.61	1200	0.68
adj. M	-1199.51	1200	1.00

Nuevamente, la prueba no es significativa, es decir, la estructura de covarianzas no es la misma.

5.1.3.2. Pruebas T^2 de Hotelling para diferencias entre medias.

	Estadístico F	Df1	Df2	P-value
G1-G2	2.5566	24	4	0.1871
G1-G3	4.00	24	4	0.09317
G1-G4	4.3068	24	4	0.0824
G1-G5	1.0329	24	4	0.5569
G2-G3	3.7946	24	4	0.1015
G2-G4	8.3731	24	4	0.02575
G2-G5	3.1342	24	4	0.1376
G3-G4	11.078	24	4	0.01537
G3-G5	3.9389	24	4	0.09556
G4-G5	3.6373	24	4	0.1087

En este caso, al nivel $\alpha = 0.05$ la mayoría de las pruebas no son significativas[3]. Sin embargo, al nivel $\alpha = 0.01$ solo en la primera (G1-G2), cuarta (G1-G5) y séptima (G2-G5) pruebas los vectores de medias son estadísticamente iguales (pruebas no significativas). En la siguiente tabla se observa que estas diferencias no significativas corresponden a las mínimas distancias de Mahalanobis.

[3] Únicamente dos de las pruebas son significativas, específicamente entre los Grupos 2 y 4, y los Grupos 3 y 4

5.1.3.3. Distancias de Mahalanobis entre regiones.

	G1	G2	G3	G4	G5
G1	0.00	100.586	194.42	209.31	50.2
G2		0.00	210.767	465.065	174.086
G3			0.00	717.909	255.2451
G4				0.00	235.6987
G5					0.00

5.1.3.4. Tabla de clasificación *plug-in* para discriminación de estados.

	G1	G2	G3	G4	G5	Error
G1	10	0	0	0	0	0.000
G2	0	7	0	0	0	0.000
G3	0	0	5	0	0	0.000
G4	0	0	0	5	0	0.000
G5	0	0	0	0	5	0.000

Los resultados del Análisis Discriminante apoyan las regiones propuestas para el caso cuando se consideran todos los indicadores. Ésto puede corroborarse con la tabla anterior, donde todos los estados se clasificaron en la región propuesta.

5.1.4. Resultados con el promedio de los indicadores.

A continuación se presentan los resultados del Análisis Discriminante con el promedio de los indicadores de la VNM2005 y la VNM2006.

5.1.4.1. Prueba de homogeneidad en la estructura de covarianzas.

	Estadístico	Df	P-value
Box. M	220.1321	312	0.99
adj. M	-5.651	312	1.00

Nuevamente, la prueba no es significativa, es decir, la estructura de covarianzas no es la misma, pero se recuerda que en este proyecto se trabajó bajo el supuesto de homogeneidad en la estructura de covarianzas.

5.1.4.2. Pruebas T^2 de Hotelling para diferencias entre medias.

	Estadístico F	Df1	Df2	P-value
G1-G2	5.4112	12	16	0.0011
G1-G3	7.2718	12	16	0.0002
G1-G4	9.2885	12	16	0.0000
G1-G5	2.3754	12	16	0.05394
G2-G3	1.8501	12	16	0.1245
G2-G4	18.569	12	16	0.0000
G2-G5	6.8712	12	16	0.0002
G3-G4	20.875	12	16	0.0000
G3-G5	9.692	12	16	0.0000
G4-G5	3.613	12	16	0.0092

La única diferencia no significativa se presenta entre los Grupos dos y tres[4], y en la siguiente tabla se puede verificar que ésta corresponde a la mínima distancia de Mahalanobis. Las restantes diferencias sí son estadísticamente significativas.

[4]La diferencia entre los Grupos 1 y 5 no es contundente, tal como lo indica el p-value, que es ligeramente mayor a 0.05.

5.1.4.3. Distancias de Mahalanobis entre regiones.

	G1	G2	G3	G4	G5
G1	0.00	26.6116	44.1766	56.428	14.43
G2		0.00	12.8454	128.926	47.706
G3			0.00	169.088	78.5053
G4				0.00	29.2656
G5					0.00

5.1.4.4. Tabla de clasificación *plug-in* para discriminación de estados.

	G1	G2	G3	G4	G5	Error
G1	10	0	0	0	0	0.000
G2	0	7	0	0	0	0.000
G3	0	0	5	0	0	0.000
G4	0	0	0	5	0	0.000
G5	0	0	0	0	5	0.000

Nuevamente los resultados del Análisis Discriminante apoyan las regiones propuestas para el caso cuando se consideran el promedio de los indicadores. Ésto puede corroborarse con la tabla anterior, donde todos los estados se clasificaron en la región propuesta.

5.1.5. Resultados con la ponderación $(1/4, 3/4)$.

A continuación se indicadores los resultados del Análisis Discriminante con los indicadores ponderados: $\omega^{i,j} = \frac{1}{4} \cdot \omega^{i,j}_{2005} + \frac{3}{4} \cdot \omega^{i,j}_{2006}$.

5.1.5.1. Prueba de homogeneidad en la estructura de covarianzas.

	Estadístico	Df	P-value
Box. M	217.4978	312	0.9998
adj. M	-5.5833	312	1.00

Siguiendo el *patrón* que se venía presentando, la prueba no es significativa, por ende, la estructura de covarianzas estadísticamente no es la misma entre Grupos.

5.1.5.2. Pruebas T^2 de Hotelling para diferencias entre medias.

	Estadístico F	Df1	Df2	P-value
G1-G2	5.9888	12	16	0.0006
G1-G3	8.3784	12	16	0.0000
G1-G4	11.956	12	16	0.0000
G1-G5	2.3369	12	16	0.05726
G2-G3	1.328	12	16	0.2930
G2-G4	23.816	12	16	0.0000
G2-G5	8.375	12	16	0.0000
G3-G4	26.042	12	16	0.0000
G3-G5	10.964	12	16	0.0000
G4-G5	4.6143	12	16	0.0027

De igual manera que en la sección anterior, la única diferencia no significativa se presenta entre los Grupos dos y tres[5].

[5] Nuevamente la diferencia entre los Grupos 1 y 5 no es contundente, tal como lo indica el p-value, que es ligeramente mayor a 0.05.

5.1.5.3. Distancias de Mahalanobis entre regiones.

	G1	G2	G3	G4	G5
G1	0.00	29.4525	50.8992	72.633	14.197
G2		0.00	9.2207	165.357	58.147
G3			0.00	210.9423	88.8102
G4				0.00	37.3765
G5					0.00

5.1.5.4. Tabla de clasificación *plug-in* para discriminación de estados.

	G1	G2	G3	G4	G5	Error
G1	10	0	0	0	0	0.000
G2	0	6	1	0	0	0.1428
G3	0	1	5	0	0	0.2000
G4	0	0	0	5	0	0.0000
G5	0	0	0	0	5	0.0000

Los resultados son similares a los obtenidos con la VNM2006 (sección 5.1.2). Hay un estado propuesto para la Región 2 que en el Análisis Discriminante se clasificó en el Grupo 3: Querétaro; y un estado propuesto en la Región 3, que se clasifica en el Grupo 2: Puebla. De alguna manera estas similitudes son lógicas pues en esta sección se le está dando mayor *peso* a la Verificación del 2006 (3/4). Nuevamente, podemos interpretar que la clasificación con el Análisis Discriminante conincide con la regionalización propuesta.

5.2. Análisis Discriminante para la VNM2003.

En esta sección el objetivo es presentar los resultados del Análisis Discriminante para los datos de la VNM2003 y verificar si las regiones utilizadas en ésta coinciden con las regiones finales propuestas con base a indicadores electorales. Para mantener la estructura y con el objeto de comparar resultados, se analizarán los mismos indicadores que se han manejado en todo el proyecto:

- **Encuesta de Cobertura.**
  1. Empadronados.
  2. Empadronados en el estado.
  3. Empadronados en la sección.
  4. Credencializados.
  5. Credencializados en el estado.
  6. Credencializados en la sección.
- **Encuesta de Actualización.**
  1. Fallecidos en el padrón.
  2. Cambios de domicilio no reportados en el padrón.
  3. Cambios de domicilio no reportados al mismo municipio.
  4. Cambios de domicilio no reportados a otro municipio dentro del mismo estado.
  5. Cambios de domicilio no reportados a otro estado.
  6. Cambios de domicilio no reportados a otro país.

Fue precisamente en esta Verificación (2003) la última donde se realizaron inferencias a nivel regional. En efecto, con el objeto de conocer con mayor detalle

geográfico la evolución de la calidad del padrón se utilizaron las 10 regiones definidas en la VNM2000. Dichas regiones se definieron en función de los trabajos sobre "marginación" realizados por el Consejo Nacional de Población (CONAPO) y el supuesto principal es que el comportamiento de cada estado era similar al de la región a la que pertenecía. Las "*ventajas*" de esta regionalización son, por un lado, que agrupa estados de la frontera norte que tienen características peculiares por la migración internacional, y por otro, agrupa estados de la zona sur que tienen una importante presencia de población indígena. Sin embargo, como se ha mencionado, la conformación de las regiones de CONAPO se basan en el Índice de Marginación y no en información sobre el comportamiento electoral del país.

5.2.1. Resultados.

La siguiente es la tabla de clasificación *plug-in* para discriminación de estados.

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
G1	3	0	0	0	0	0	0	0	0	0
G2	0	1	0	0	0	0	0	1	0	0
G3	0	0	4	0	0	0	0	0	0	0
G4	0	0	0	3	0	0	0	0	0	0
G5	0	0	1	0	2	0	0	0	0	0
G6	0	0	0	0	0	3	0	0	0	0
G7	0	0	0	0	0	0	3	0	0	0
G8	0	0	0	0	0	0	0	4	0	0
G9	0	0	0	0	0	1	0	0	2	0
G10	0	0	0	0	0	0	0	0	0	4

La siguiente es la tabla de validación (cross-validation table) para la discriminación de estados [6].

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
G1	0	0	1	1	0	0	0	0	0	1
G2	0	0	0	0	1	0	0	1	0	0
G3	0	0	0	0	1	1	1	0	1	0
G4	1	0	0	0	0	0	0	1	1	0
G5	0	1	1	0	0	0	0	1	0	0
G6	0	0	0	0	1	0	0	0	2	0
G7	1	1	0	0	0	0	0	1	0	0
G8	0	1	0	0	0	0	0	2	0	1
G9	0	0	0	0	0	1	0	0	2	0
G10	0	0	0	0	2	0	0	0	0	1

Comparando estos resultados con aquellos obtenidos en la sección anterior con el Análisis Discriminante para cada modalidad, se observa que la propuesta de diez regiones de CONAPO no presenta tan buen comportamiento interno en la tabla de validación. La tabla de clasificación *plug-in* sí presenta cierta consistencia en la conformación de regiones, más sin embargo, dos ventajas de las regiones propuestas en este proyecto son:

1. Las regiones propuestas son cinco,
2. Los resultados de los análisis arrojan que la consistencia en estas cinco regiones es mejor que la propuesta de CONAPO.

Es decir, las regiones propuestas se reducen a la mitad y más aún, se presenta un mejor comportamiento.

[6] En la tabla de *validación cruzada* (cross-validation table), un *individuo* se clasifica de acuerdo con la función discriminante derivada de todos los *individuos* restantes, eliminando así al primero.

5.3. Estadísticos descriptivos de las regiones propuestas.

En esta sección se exhiben una serie de medidas descriptivas de las cinco regiones finales propuestas en el capítulo anterior, el objetivo es tener un panorama general de éstas. Asimismo, se presentarán Gráficas de Caja (Box-Plot) para los siguientes indicadores: Empadronados, Credencializados, Fallecidos, Cambios de domicilio no reportados al Padrón, Cambios de domicilio a otro estado no reportados y Cambios de domicilio a otro país no reportados. Finalmente se exhibirá un mapa de la República Mexicana con las regiones con el objeto de ubicarlas geográficamente.

Con base en los resultados de capítulos previos, se determinó presentar las medidas descriptivas para el caso de los indicadores ponderados $(1/4, 3/4)$, es decir, los indicadores definidos con la siguiente combinación lineal:

$$\omega^{i,j} = \frac{1}{4}\omega^{i,j}_{2005} + \frac{3}{4}\omega^{i,j}_{2006}$$

pues como se pudo ver en los distintos análisis, la VNM2006 tuvo un mayor *peso* en la conformación de las diferentes regiones.

5.3.1. Tablas de estadísticos descriptivos.

Por cada región final se reportan los siguientes estadísticos: Promedio, Desviación Estándar, Mínimo, Máximo, Sesgo y Curtosis, de cada indicador. Debe recordarse que los indicadores están reportados en términos de porcentajes por lo cual, estos estadísticos representan dichos porcentajes.

REGIÓN 1

	Promedio	Std. Dev.	Mínimo	Máximo	Sesgo	Curtosis
Empadronados	95.133	0.91456	93.417	96.59	-0.1826	2.6558
Empadronados en el estado.	93.1725	0.4463	92.28	93.83	-0.6210	2.7486
Empadronados en la sección.	77.027	2.16	73.40	80.85	-0.0542	2.621
Credencializados.	91.95	1.355	90.085	93.795	-0.0377	1.633
Credencializados en el estado.	87.6105	1.0617	85.7	89.38	-0.094	2.598
Credencializados en la sección.	75.46	2.1026	72.14	79.407	0.0041	2.8524
Fallecidos.	1.6145	0.83	0.78	3.332	0.985	2.911
Cambios de domicilio no reportados en el padrón.	22.20	2.111	19.327	24.962	-0.0594	1.3707
Cambios de domicilio al mismo municipio.	9.93	1.968	7.37	13.19	0.3148	1.8681
Cambios de domicilio a otro municipio dentro del mismo estado.	2.101	0.7056	0.895	3.27	-0.055	2.288
Cambios de domicilio a otro estado.	1.971	0.476	1.2425	2.61	-0.0896	1.90
Cambios de domicilio a otro país.	5.817	1.981	2.87	8.937	0.23	2.072

REGIÓN 2

	Promedio	Std. Dev.	Mínimo	Máximo	Sesgo	Curtosis
Empadronados	94.86	0.651	94.045	95.96	0.5051	2.226
Empadronados en el estado.	92.41	1.3773	90.635	94.56	0.2405	2.00
Empadronados en la sección.	78.31	1.797	75.82	80.44	-0.0636	1.591
Credencializados.	91.36	1.3812	89.187	93.115	-0.2668	1.927
Credencializados en el estado.	87.601	2.2024	84.56	91.09	0.3503	2.1686
Credencializados en la sección.	76.30	1.911	73.737	78.687	-0.358	1.672
Fallecidos.	1.2164	0.3938	0.495	1.6575	-0.8050	2.607
Cambios de domicilio no reportados en el padrón.	16.946	1.5153	15.502	19.087	0.4587	1.519
Cambios de domicilio al mismo municipio.	7.74	1.222	5.57	8.80	-1.007	2.331
Cambios de domicilio a otro municipio dentro del mismo estado.	2.4	0.796	1.12	3.27	-0.4184	1.889
Cambios de domicilio a otro estado.	2.179	0.512	1.375	2.825	-0.1458	2.034
Cambios de domicilio a otro país.	2.03	0.7937	0.995	3.44	0.5072	2.6168

REGIÓN 3

	Promedio	Std. Dev.	Mínimo	Máximo	Sesgo	Curtosis
Empadronados	96.723	1.4565	94.36	98.005	-0.887	2.437
Empadronados en el estado.	94.48	1.9337	92.36	96.75	-0.135	1.386
Empadronados en la sección.	82.4	1.433	81.44	84.93	1.426	3.16
Credencializados.	93.73	1.68	91.115	95.7	-0.5941	2.441
Credencializados en el estado.	90.04	2.62	87.75	93.95	0.6382	1.864
Credencializados en la sección.	80.91	1.515	79.81	83.54	1.30	2.97
Fallecidos.	1.24	0.4711	0.5125	1.7975	-0.5498	2.402
Cambios de domicilio no reportados en el padrón.	16.77	1.495	14.92	18.66	-0.0617	1.65
Cambios de domicilio al mismo municipio.	6.49	2.096	3.87	8.762	-0.0922	1.456
Cambios de domicilio a otro municipio dentro del mismo estado.	2.34	0.4061	1.967	3.017	1.0378	2.66
Cambios de domicilio a otro estado.	2.296	0.864	1.475	3.525	0.467	1.7067
Cambios de domicilio a otro país.	2.093	0.7218	1.31	3.03	0.321	1.505

REGIÓN 4

	Promedio	Std. Dev.	Mínimo	Máximo	Sesgo	Curtosis
Empadronados	94.27	1.291	92.82	95.47	-0.2558	1.234
Empadronados en el estado.	88.74	3.65	83.85	93.42	-0.0333	1.9090
Empadronados en la sección.	64.88	5.608	55.39	69.46	-1.126	2.7552
Credencializados.	87.75	3.374	84.11	92.09	0.3051	1.4642
Credencializados en el estado.	80.37	4.148	74.195	84.46	-0.4886	1.996
Credencializados en la sección.	62.28	6.306	51.47	67.71	-1.187	2.888
Fallecidos.	1.418	0.7263	0.70	2.597	0.86	2.517
Cambios de domicilio no reportados en el padrón.	29.16	2.362	25.92	32.582	0.132	2.471
Cambios de domicilio al mismo municipio.	13.95	1.487	12.44	15.515	-0.0352	1.2593
Cambios de domicilio a otro municipio dentro del mismo estado.	2.76	1.04	1.3325	4.13	-0.08114	2.0775
Cambios de domicilio a otro estado.	2.632	1.465	1.2025	4.73	0.368	1.82
Cambios de domicilio a otro país.	3.735	2.787	0.8	7.292	0.1169	1.488

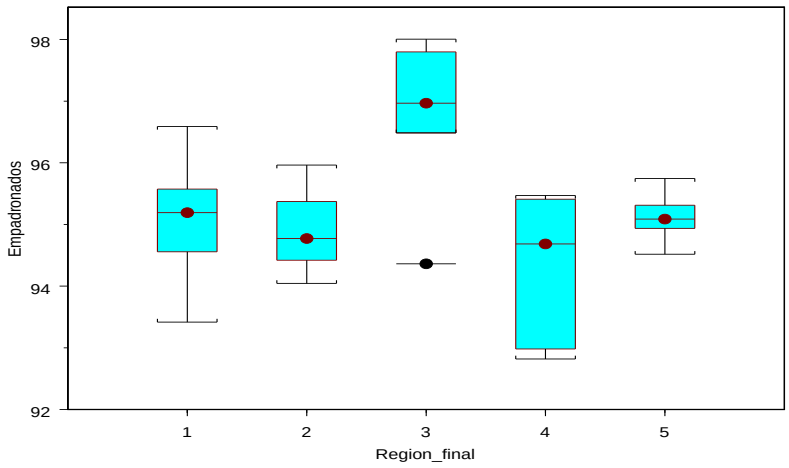
REGIÓN 5

	Promedio	Std. Dev.	Mínimo	Máximo	Sesgo	Curtosis
Empadronados	95.12	0.454	94.51	95.74	0.0789	2.107
Empadronados en el estado.	92.55	1.02	91.727	94.27	1.093	2.71
Empadronados en la sección.	71.57	1.2923	70.38	73.38	0.50	1.623
Credencializados.	91.82	0.4234	91.46	92.54	1.07	2.71
Credencializados en el estado.	86.68	1.546	85.23	88.82	0.444	1.585
Credencializados en la sección.	69.87	1.08	69.00	71.36	0.5113	1.493
Fallecidos.	1.368	0.2713	1.05	1.76	0.318	2.015
Cambios de domicilio no reportados en el padrón.	24.57	1.53	22.11	26.14	-0.8248	2.446
Cambios de domicilio al mismo municipio.	11.77	1.323	10.00	13.29	-0.2113	1.673
Cambios de domicilio a otro municipio dentro del mismo estado.	3.94	2.223	2.0275	7.48	0.8477	2.267
Cambios de domicilio a otro estado.	1.8525	0.7124	1.2	2.92	0.6175	1.9472
Cambios de domicilio a otro país.	3.981	1.152	2.432	5.15	-0.4085	1.5115

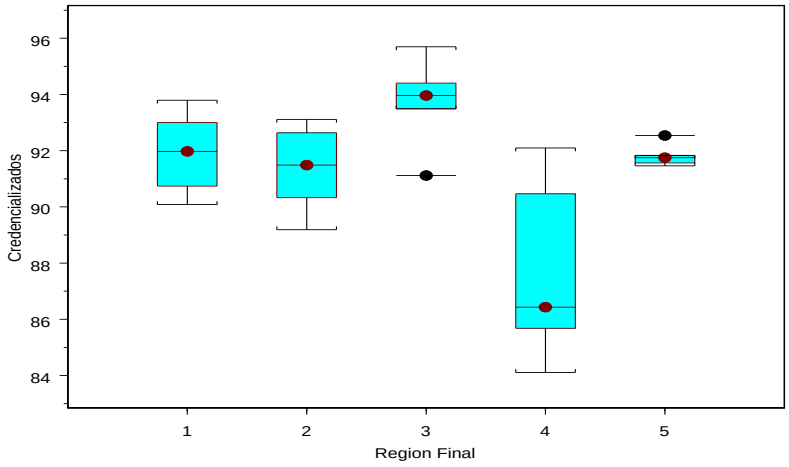
5.3.2. Gráficas de Caja (Box-Plot).

A continuación se exhiben las Gráficas de Caja que permiten verificar gráficamente el comportamiento de los siguientes indicadores: Empadronados, Credencializados, Fallecidos, Cambios de domicilio no reportados al Padrón, Cambios de domicilio a otro estado no reportados y Cambios de domicilio a otro país no reportados. Se presentan por región final para que sean fácilmente comparables.

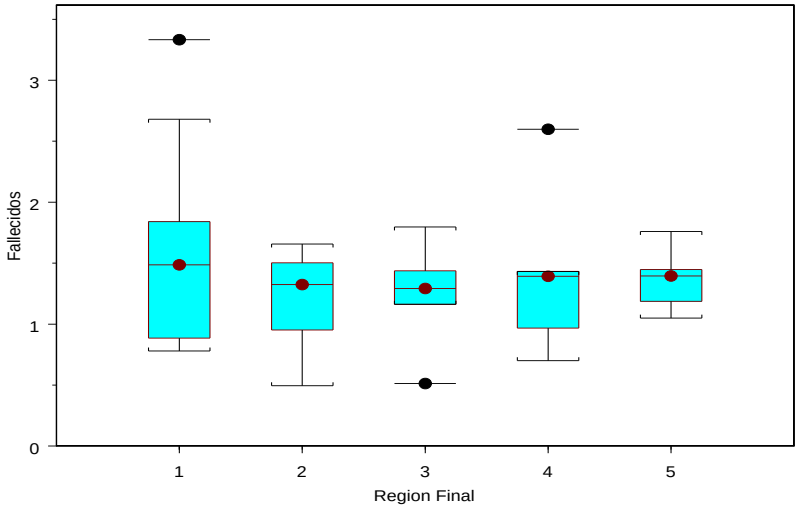
Región Final vs. Empadronados



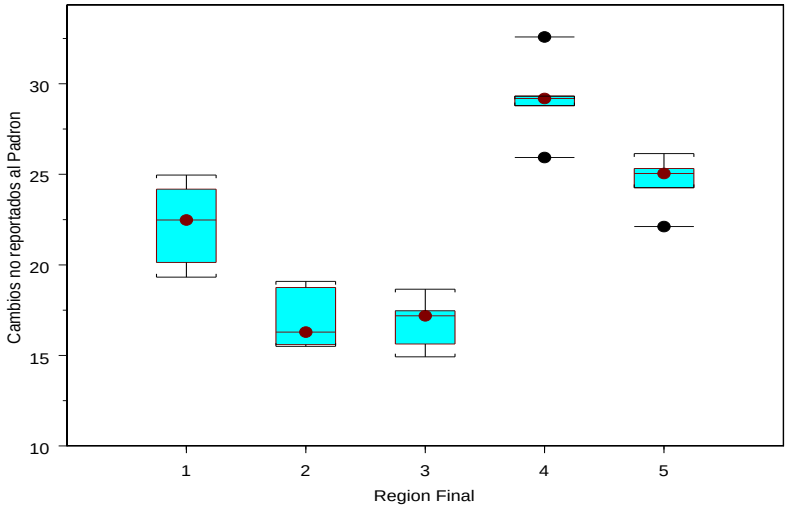
Región final vs. Credencializados



Región final vs. Fallecidos

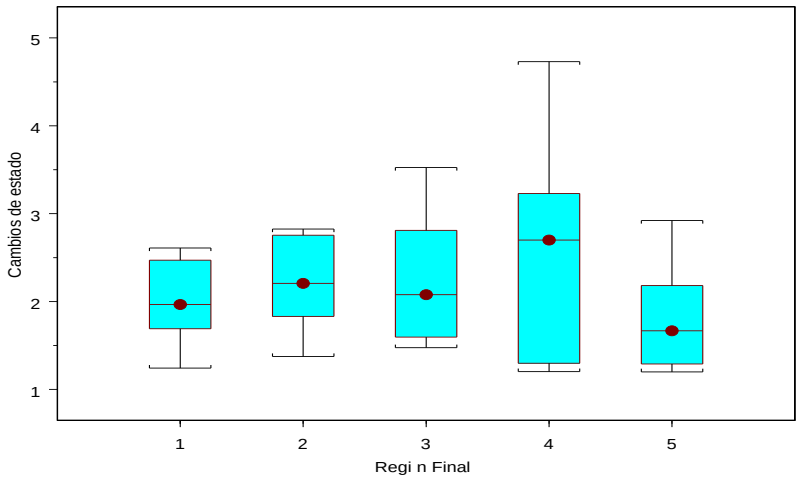


Región final vs. Cambios de domicilio no reportados al Padrón.

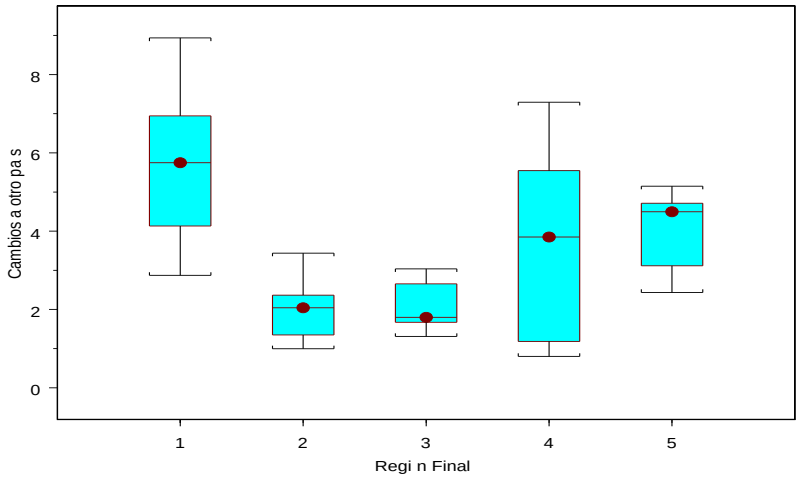


Pueden observarse algunos valores *atípicos*, principalmente con el indicador *Fallecidos*. Y la región que ha presentado menor variabilidad es la Región 5.

Región final vs. Cambios de domicilio a otro estado no reportados.



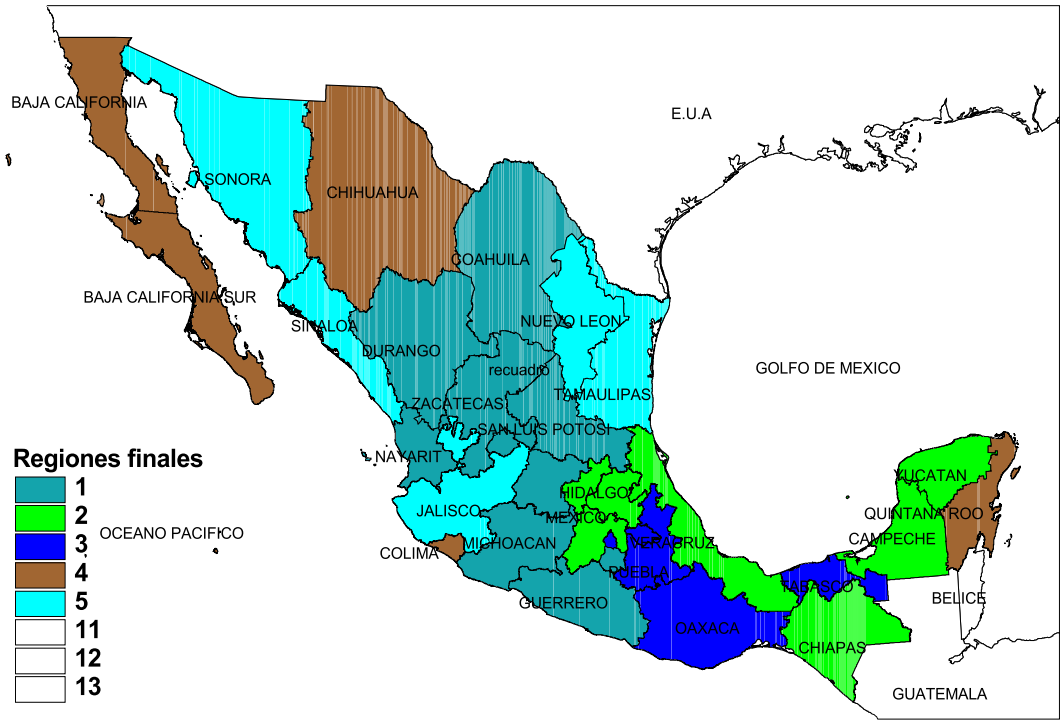
Región final vs. Cambios de domicilio a otro país no reportados.



Nótese que la Región 4 es la que presenta la mayor variabilidad mientras que la Región 5 presenta la menor, tal como lo reflejaron las tablas con las medidas descriptivas previamente expuestas. De los indicadores presentados, los que más semejanza guardan en cuanto a estimación puntual entre regiones finales son Fallecidos y Cambios de domicilio a otro estados no reportados.

A continuación se muestra el mapa de la República Mexicana con las regiones finales propuestas.

Región 1	Aguascalientes, Coahuila, Guanajuato, Nayarit, Zacatecas, Michoacán, San Luis Potosí, Morelos, Durango, Guerrero.
Región 2	Campeche, Chiapas, Querétaro, Veracruz, México, Yucatán, Hidalgo.
Región 3	Distrito Federal, Tlaxcala, Tabasco, Puebla, Oaxaca.
Región 4	Baja California, Chihuahua, Colima, Quintana Roo, Baja California Sur.
Región 5	Sonora, Sinaloa, Nuevo León, Jalisco, Tamaulipas.



5.3.3. Conclusiones.

En el capítulo anterior, se realizaron análisis en distintas modalidades con las Verificaciones Nacionales Muestrales en donde, *a priori*, se pudo verificar que la VNM2006 tiene un mayor impacto en las regiones propuestas. Con el Análisis Discriminante de este capítulo se corrobora lo anterior y se tienen elementos suficientes para validar las regiones propuestas.

Este comportamiento validado de las entidades federativas representa una regionalización alterna y completamente diferente a la propuesta de CONAPO con base en el Índice de Marginación. Ahora bien, cabe aclarar que las 10 regiones utilizadas en la VNM2003 (Propuestas por CONAPO y cuyo Análisis Discriminante se exhibió en la sección 5.2), y las regiones propuestas en este proyecto, no son comparables indicador a indicador debido a que las regiones de este proyecto se realizaron con información a nivel estatal de las Verificaciones 2005 y 2006. En un futuro, si se proponen nuevamente regiones con base a información estatal de las Verificaciones Nacionales Muestrales [7], entonces será posible realizar comparaciones (Pruebas de hipótesis para diferencia entre indicadores, etc.)

Los resultados obtenidos con este proyecto no pretenden desacreditar la propuesta de CONAPO, sino por el contrario, el objetivo es proponer una regionalización de los Estados Unidos Mexicanos basada en índices electorales recientes que sea útil para realizar las inferencias adecuadas. Los resultados de este proyecto sugieren que, para futuras Verificaciones, se considere el hecho de que ambos comportamientos presentan diferencias marcadas en cuanto a las regiones que presentan. Asimismo, es importante señalar que debido a distintos factores de-

[7] Como por ejemplo, realizar una regionalización para la Verificación del año 2008.

mográficos, sociales, etc., se piensa que las regiones pueden variar en el tiempo, en ese sentido es de interés realizar estos ejercicios de regionalización con base a indicadores electorales previo al levantamiento de las Verificaciones, con el objeto de poder realizar inferencias a nivel regional y que la estimación de los principales indicadores sea con mayor precisión.

Apéndice A

El Índice de Gini.

En las distribuciones de variables económicas, como son la producción, los salarios, o las rentas, el interés se centra en el conocimiento del reparto equitativo de sus valores. Así, cuando se estudia la nómina de una empresa, interesa conocer si está bien repartida entre todos sus trabajadores o se encuentra concentrada en unos pocos. Corrado Gini, con la intención de medir el grado de equidistribución de una variable, introduce el concepto de *concentración*. Se entiende por *concentración* el mayor o menor grado de igualdad en el reparto del total de los valores de la variable.

Consideremos la distribución dada por los valores de la variable acompañados de los valores de sus respectivas frecuencias, (x_i, n_i), con $i = 1, \ldots, k$, donde los valores x_i están ordenados de menor a mayor, y $x_i \geq 0$ para toda i. Su concentración puede ser estudiada gráficamente o a través de un Índice. En cualquier caso y para un mejor entendimiento del Índice de Gini, es necesario introducir unos conceptos previos y así, finalmente, exhibir su definición:

Se llama *masa parcial* correspondiente a un valor x_i de una variable X, al resultado de multiplicar el valor de la variable por su frecuencia absoluta: $x_i \cdot n_i$.

Se denomina *masa parcial acumulada* hasta un determinado valor de la variable, a la suma de las masas parciales de los valores de la variable menores o iguales a él. Es decir, si se denota la *masa parcial acumulada* como U_i, entonces:

$$U_i = \sum_{j=1}^{n_i} x_j n_j \qquad i = 1, \ldots, k.$$

Se llama *masa total de la variable* (M) a la suma de todas sus masa parciales:

$$M = \sum_{i=1}^{k} x_i \cdot n_i$$

Además:

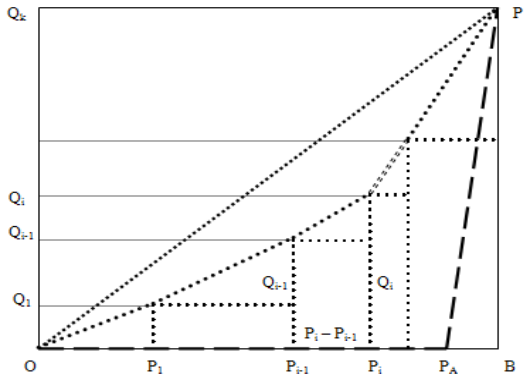
1. $P_i = 100 \times F_i = 100\frac{N_i}{N}$, es la frecuencia absoluta acumulada en x_i en forma de porcentaje, siendo $N_i = \sum_{j=1}^{n} n_j$ la frecuencia absoluta acumulada en x_i.
2. $Q_i = 100 \times \frac{U_i}{M}$ es la masa parcial acumulada en x_i expresada en porcentaje.
3. $D_i = 100 \times \frac{x_i \cdot n_i}{M}$ es el porcentaje de la masa parcial que perciben los n_i individuos que forman parte de la i-ésima clase.
4. $C_i = 100 \times \frac{n_i}{N}$ es el porcentaje de individuos que perciben la masa parcial $x_i \cdot n_i$.

A.1. Estudio gráfico. La Curva de Lorenz.

El estudio gráfico de este problema se realiza por medio de la *curva de concentración o curva de Lorenz*, que es la representación de las masas parciales

acumuladas en porcentajes (Q_i) en función de las frecuencias acumuladas en porcentajes (P_i).

Para dibujar la curva de Lorenz[1], se suele construir un cuadrado de lado 100 unidades, tomando como origen el vértice inferior izquierdo, situando sobre el eje de abcisas las P_i, y sobre el eje de las ordenadas las Q_i. La poligonal que une los puntos (P_i, Q_i) es la *curva de Lorenz* o *curva de concentración*. Pasa por los puntos $O(0,0)$ y $P(100,100)$, y está situada siempre por debajo de la diagonal que une O y P, siendo convexa, ya que $Q_i \leq P_i$, para $i = 1, \ldots, k$.



La diagonal OP del cuadrado, que une los puntos $O(0,0)$ y $P(100,100)$ se denomina *recta de equidistribución*. Se llama región de concentración a la superficie comprendida entre la recta de equidistribución y la curva de concentración. La medida de esta superficie se denomina *área de concentración*(AC). En caso de que se diera la máxima concentración en el reparto de los valores de una variable económica, como puede ser una nómina, sucedería que, para una población de N empleados, $N-1$ de los empleados no tendrían salario y el otro empleado se llevaría toda la masa salarial de modo que la correspondiente curva de Lorenz

[1] al trabajar con porcentajes.

sería la curva de *máxima concentración*, y vendría dada por OP_AP, siendo:

$$P_A = \frac{N-1}{N} \times 100$$

Ahora bien, cuando el número de individuos es muy grande (lo cual se admite si $N > 100$), en una situación de máxima concentración, P_A tiende a $B(100, 0)$, y, en tal caso, la curva de máxima concentración estará formada por los lados inferior y lateral derecho del cuadrado. El área comprendida entre la recta de equidistribución y la curva de máxima concentración se denomina *área de máxima concentración* (AMC). La concentración mínima se daría cuando todos los empleados percibieran el mismo salario, en cuyo caso $P_i = Q_i$ para toda i, y la curva de Lorenz coincidiría con la diagonal del cuadrado. Por lo tanto, cuanto más se aproxime la curva de Lorenz a la diagonal del cuadrado, menor será la concentración y más equitativa será la distribución de los salarios.

A.2. Índices de Concentración basados en la Curva de Lorenz.

En general, son diversos los Índices que han sido propuesto y que tratan de dar una medida de la *concentración*[2] , los cuales tienen distintas aplicaciones. En particular, tres de ellos tienen su fundamento en la curva de Lorenz: el *Índice geométrico (RC)*, el *Índice asintótico (I)* y el *Índice de Gini*, (I_G).

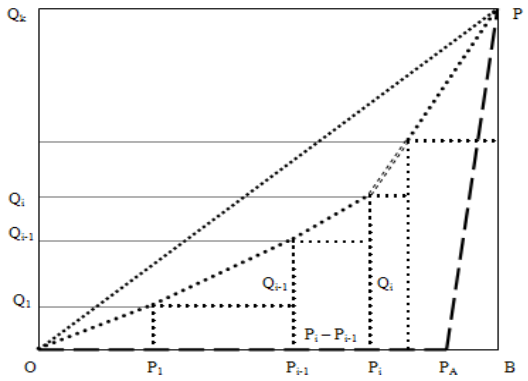
A.2.1. Él Indice geométrico o Razón de Concentraciones(RC).

Cuanto mayor sea el área de concentración respecto del área de máxima concentración, mayor será la desigualdad en el reparto de la masa total de la variable.

[2] Concepto definido previamente.

Por ello, el cociente entre el área de concentración y el área de máxima concentración es también una medida de concentración que se conoce como *Índice geométrico o Razón de Concentraciones (RC).*

El área de concentración se puede obtener restando el área del triángulo (mitad del área del cuadrado de lado 100), cuya medida es 5,000 unidades cuadradas, la suma de las áreas de los recintos que quedan por debajo de la curva de concentración.



Estos recintos son trapecios de altura $P_i - P_{i-1}$ y bases Q_{i-1} y Q_i. Entonces el área de concentración viende dada por:

$$AC = \frac{10,000}{2} - \frac{\sum_{i=1}^{k} (P_i - P_{i-1})(Q_i + Q_{i-1})}{2}$$

El área de máxima concentración es:

$$AMC = 100 \times \frac{P_A}{2}$$

siendo P_A el valor de P_i cuando se supone que $N-1$ elementos de la distribución toman el valor cero y el elemento restante toma un valor igual a la masa total de la variable:

$$P_A = \frac{N-1}{N} \times 100$$

Por lo tanto, el Índice geométrico es:

$$\begin{aligned} RC &= \frac{AC}{AMC} = \frac{10,000 - \sum_{i=1}^{k} (P_i - P_{i-1})(Q_i + Q_{i-1})}{100 P_A} \\ &= \frac{N}{N-1}\left[1 - \frac{\sum_{i=1}^{k} (P_i - P_{i-1})(Q_i + Q_{i-1})}{10,000}\right] \end{aligned}$$

El Índice geométrico se interpreta en el siguiente sentido:

1. Cuando $RC \approx 0$, el área de concentración vale cero, lo que sucede si $P_i = Q_i$, para $i = 1, 2, \ldots, k$. Es decir, en la medida en que la *Razón de Concentraciones* (RC) se aproxime a cero, habrá una mayor equidistribución en el reparto de la masa total de la variable.

2. Cuando $RC = 1$, entonces AC y AMC coinciden, lo cual significa que la concetración es máxima, es decir, un solo elemento de la distribución acapara toda la masa de la variable.

3. Cuanto más se aleja RC de cero y se aproxima a uno, tanto mayor es el nivel de concentración de la masa total de la variable.

A.2.2. El Índice asintótico.

Se toma como Índice de concentración asintótico (I), el doble del área de la región de concentración cuando se trabaja sobre el cuadrado de lado unidad, precisamente porque a este valor converge asintóticamente el Índice geométrico. De esta forma, el Índice de concentración asintótico, cuando se considera el cuadrado de lado 100, es:

$$I = \frac{2 \cdot AC}{10,000}$$

Este Índice varía de cero a uno. Vale cero cuando la curva de concentración coincide con la recta de equidistribución, y vale uno cuando la curva de concentración coincide con los lados del cuadrado.

A.2.3. El Índice de Gini.

Para distribuciones contínuas, Gini define el Índice de concentración como el doble del área de concentración, es decir, a través del Índice asintótico (I). Sin embargo, para obtener una evaluación del Índice de concentración, se sirve de una variable discreta con frecuencias unitarias, obteniendo la siguiente expresión:

$$I_G = \frac{\sum_{i=1}^{N-1}(P_i - Q_i)}{\sum_{i=1}^{N-1} P_i} = 1 - \frac{\sum_{i=1}^{N-1} Q_i}{\sum_{i=1}^{N-1} P_i}$$

para $i = 1, 2, \ldots, N-1$, ya que $P_N = Q_N$. donde N es el tamaño de la población.

A.2.4. El Índice generalizado de Gini.

Esta expresión, válida para frecuencias unitarias, se ha generalizado y aparece en diversos textos y/o artículos aplicada a distribuciones con frecuncias no unitarias en la formas[3]:

$$I_G^{'} = \frac{\sum_{i=1}^{k-1}(P_i - Q_i)}{\sum_{i=1}^{k-1} P_i} = 1 - \frac{\sum_{i=1}^{k-1} Q_i}{\sum_{i=1}^{k-1} P_i}$$

donde k representa el número de clases diferentes, siendo $k < N$. Pero esta expresión del Índice de Gini para frecuencias no unitarias no es correcta, por lo que su aplicación para calcular el grado de concentración de una variable no es válida, proporcionando, en algunos casos, resultados contradictorios. Para resolver este problema, se debe expresar el Índice geométrico en función de las frecuencias[2]. Considérense entonces los siguientes resultados:

Proposición. *Sea F la distribución de la variable no negativa X dada por los pares (x_i, n_i), $i = 1, 2, \ldots, k$, siendo k el número de clases; x_1, x_2, ..., x_k los*

[3] Vargas, S., "Expresión del Índice de Gini para frecuencias no unitarias."

valores de la variable ordenados en orden no decreciente, y n_1, n_2, ..., n_k *las frecuencias absolutas respectivas de cada una de las clases. Entonces se verifica:*

$$\sum_{i=1}^{k} (P_i - P_{i-1})(Q_i + Q_{i-1}) = \frac{10,000}{MN}\left[2\sum_{i=1}^{k} n_i U_i - \sum_{i=1}^{k} n_i^2 x_i\right]$$

Prueba. De la definición de P_i y C_i, así como de Q_i y D_i, resulta:

$$\begin{aligned}
\sum_{i=1}^{k} (P_i - P_{i-1})(Q_i + Q_{i-1}) &= \sum_{i=1}^{k} (P_i - P_i + C_i)(Q_i + Q_i - D_i) \\
&= \sum_{i=1}^{k} (C_i)(2Q_i - D_i) \\
&= 2\sum_{i=1}^{k} C_i Q_i - \sum_{i=1}^{k} C_i D_i \\
&= \frac{2 \times 100}{N}\sum_{i=1}^{k} n_i Q_i - \frac{100}{N}\sum_{i=1}^{k} n_i D_i \\
&= \frac{200}{N}\left[\sum_{i=1}^{k} n_i U_i \frac{100}{M} - \frac{1}{2}\sum_{i=1}^{k} n_i (x_i n_i)\frac{100}{M}\right] \\
&= \frac{10,000}{N \cdot M}\left[2\sum_{i=1}^{k} n_i U_i - \sum_{i=1}^{k} n_1^2 x_i\right]
\end{aligned}$$

■

De la proposición anterior, se deducen los siguientes resultados:

Corolario 1. *Bajo las hipótesis de la Proposición 1, si se denota* $U = \sum_{i=1}^{k} n_i U_i$, *la suma ponderada de las masas parciales acumuladas, y* $W = \sum_{i=1}^{k} n_i^2 x_i$, *la masa total ponderada de la variable, se verifican:*

1. $I = 1 - \frac{2U-W}{MN}$
2. $RC = \frac{N}{N-1} \cdot I$
3. $RC = 1 - \frac{2U-W-M}{M(N-1)}$

Prueba. En efecto, el área de concentración viene dada por:

$$AC = 5,000 - \frac{5,000}{MN}\left[\sum_{i=1}^{k} 2n_i U_i - \sum_{i=1}^{k} n_i^2 x_i\right] = 5,000\left[1 - \frac{2U - W}{MN}\right]$$

y, teniendo en cuenta la definición de I, se sigue:

$$I = \frac{2 \cdot AC}{10,000} = 1 - \frac{2U - W}{MN}$$

y queda demostrado 1). Por otro lado, el índice geométrico es:

$$\begin{aligned} RC &= \frac{AC}{AMC} = \frac{5,000\left(1 - \frac{2U-W}{MN}\right)}{5,000\frac{N-1}{N}} \\ &= \frac{N}{N-1}\left(1 - \frac{2U - W}{MN}\right) \\ &= \frac{N}{N-1} I \end{aligned}$$

lo cual demuestra 2). Entonces:

$$\begin{aligned} RC &= \frac{N}{N-1} I = \frac{N-1+1}{N-1} I \\ &= 1 + \frac{1}{N-1} - \frac{2U - W}{M(N-1)} \\ &= 1 + \frac{M}{(N-1)M} - \frac{2U - W}{M(N-1)} \\ &= 1 - \frac{2U - W - M}{M(N-1)} \end{aligned}$$

lo cual prueba 3).

■

Se toma entonces como expresión generalizada del Índice de Gini y se denotará por I_v, a la expresión obtenida en 3):

$$\boxed{I_v = 1 - \frac{2U-W-M}{M(N-1)}}$$

Corolario 2. *El Índice I_v converge asintóticamente al Índice I.*

Prueba. En efecto, cuando $N \to \infty$, $\frac{N}{N-1} \to 1$, y por tanto, de la relación 2) del Corolario 1, resulta inmediatamente que I_v converge a I.

■

Proposición 2. *Cuando las frecuencias son unitarias, se verifica que:*

1. $\sum_{i=1}^{N-1} P_i = 50(N-1)$
2. $\sum_{i=1}^{N-1} Q_i = \frac{100}{N} \sum_{i=1}^{N-1} \sum_{j=1}^{i} x_j$
3. $2U - W - M = 2\sum_{i=1}^{N-1} \sum j = 1^i x_j$

Prueba. De la Proposición 2, si las frecuencias son unitarias, resulta:

$$\begin{aligned}\sum_{i=1}^{N-1} P_i &= \sum_{i=1}^{N-1} N_i \frac{100}{N} = \frac{100}{N} \sum_{i=1}^{N-1} \sum_{j=1}^{i} 1 \\ &= \frac{100}{N} \sum_{i=1}^{N-1} i = \frac{100}{N} \frac{N(N-1)}{2} \\ &= 50(N-1)\end{aligned}$$

y queda demostrado 1). Por otro lado:

$$\sum_{i=1}^{N-1} Q_i = \sum_{i=1}^{N-1} U_i \frac{100}{M} = \frac{100}{M} \sum_{i=1}^{N-1} \sum_{j=1}^{i} x_j$$

y queda demostrado 2). En cuanto a 3), si las frecuencias son unitarias, entonces $W = M$, y se sigue que:

$$\begin{aligned}2U - W - M &= 2U - M - M \\ &= 2(U - M) = 2\sum_{i=1}^{N} (U_i - x_i) \\ &= 2\sum_{i=1}^{N} \left(\sum_{j=1}^{i} x_j - x_i \right) = 2\sum_{i=1}^{N} \sum_{j=1}^{i-1} x_j \\ &= 2\sum_{i=1}^{N-1} \sum_{j=1}^{i} x_j\end{aligned}$$

■

Corolario 3. *Si las frecuencias son unitarias, entonces* $I_V = I_G$.

Prueba. En efecto, teniendo en cuenta los resultados de la Proposición 2 para frecuencias unitarias, se tiene:

$$\begin{aligned} \frac{2U - W - M}{M(N-1)} &= \frac{2\sum_{i=1}^{N-1}\sum_{j=1}^{i} x_j}{M(N-1)} \\ &= \frac{\frac{100}{M}\sum_{i=1}^{N-1}\sum_{j=1}^{i} x_j}{50(N-1)} \\ &= \frac{\sum_{i=1}^{N-1} Q_i}{\sum_{i=1}^{N-1} P_i} \end{aligned}$$

de donde resulta inmediatamente que, si las frecuencias son unitarias, entonces $I_V = I_G$.

■

A.3. Observaciones.

La expresión del Índice de Gini I_G sólo es válida para frecuencias unitarias, de modo que, para frecuencias no unitarias y si N es grande, su evaluación resulta larga y compleja. En su lugar, se puede utilizar la siguiente expresión:

$$\boxed{I_v = 1 - \frac{2U-W-M}{M\cdot(N-1)}}$$

de la que se conoce el vector (U, W, M, N) que intervienen en ella. Se trata de una expresión sencilla, fácil de evaluar, cuyo valor siempre existe, que converge asintóticamente el Índice I y que puede ser asumida como expresión generalizada del Índice de Gini para frecuencias no unitarias.

Apéndice B

La técnica de estratificación de Dalenius.

En términos prácticos, una *estratificación* es un método que consiste en clasificar o asignar elementos disponibles a grupos inicialmente no definidos. A cada grupo formado se le denomina estrato o conglomerado, de tal forma que los elementos de un conglomerado son, en cierto sentido, "similares" o "cercanos" unos a otros, a partir de ciertas características o variables de interés.

Como se ha visto, los términos de similaridad y cercanía pueden variar dependiendo de la naturaleza de los datos; esto da lugar a que existan varias técnicas de clasificación o asignación de elementos a los conglomerados, las cuales pueden ser tipificadas en jerárquicas, de partición u optimización, de densidad o modo, entre otras.

La *técnica de estratifiación de Dalenius* consiste en encontrar la mejor estratificación mediante la búsqueda de estratos cuya población sea lo más homogénea

posible, o equivalentemente, que la medida del error de la estimación o varianza de la media de cada estrato, sea mínima. Inicialmente, los datos se agrupan en una matriz X, de N renglones y K columnas, donde N es el número de unidades de que consta la población y K, el número de variables que intervienen en la estratificación. Por lo tanto, el elemento x_{ij} de la matriz X corresponde a la i-ésima observación de la j-ésima variable.

Con el objetivo de facilitar los cálculos, el valor de cada elemento de las columnas de X se transforma, entre 0 y 100. Para ello se ordenan ascendentemente los valores de cada columna y se les aplica la siguiente transformación:

$$y_{i,j} = a_j + b_j x_{i,j} \qquad i = 1, \ldots, N. \qquad j = 1, \ldots, K.$$

donde:

$$b_j = \frac{100}{max\{x_{ij}\} - min\{x_{ij}\}} \qquad \text{y} \qquad a_j = -b\left[min_i\left(x_{ij}\right)\right]$$

de tal forma que el valor mínimo de cada columna se transforma en cero, el máximo en 100 y el resto son proporcionales. Debido a esta modificación, el rango de cualquier columna es 100; este rango se divide en 10 intervalos de igual longitud y a cada columna se le aplica el siguiente procedimiento:

1. Se obtiene la frecuencia de observaciones en cada intervalo:
   $(< 0, 10], < 10, 20], \ldots, < 90, 100])$.

2. En la siguiente columna se calcula la raíz cuadrada de la frecuencia obtenida en el paso anterior.

3. Se acumulan los valores obtenidos en el punto anterior.

4. El total acumulado (T) se divide por el número de estratos que se desea formar (L).

5. Los límites óptimos de los estratos, denotados por $X^{(1)}, \ldots, X^{(L-1)}$ son:
$$X^{(1)} = \frac{T}{L}, X^{(2)} = \frac{2T}{L}, \ldots, X^{(L-1)} = \frac{(L-1)T}{L}$$
6. En la columna de valores acumulados se localizan los más cercanos entre los que queden comprendidos los límites $X^{(1)}, X^{(2)}, \ldots, X^{(L-1)}$.

Una vez definidos los estratos, se cuenta el número de elementos que tiene cada uno de ellos $(N_1, N_2, \ldots, N_L)$; generalmente estos valores difieren para cada una de las diferentes variables.

A continuación, se procede al cálculo de las varianzas, para lo cual, se supone que se obtiene una muestra de tamaño n de una población de N elementos y que los parámetros a estimar son la k medias poblacionales, digamos $\theta_1, \theta_2, \ldots, \theta_k$, del conjunto de variables transformadas $Y_1, Y_2, \ldots, Y_k$. El estimador de θ_k, usando muestreo aleatorio estratificado es:
$$\hat{\theta_k} = \frac{1}{N} \sum_{h=1}^{L} N_h \bar{Y}_{k,h} \qquad k = 1, 2, \ldots, K.$$
donde:

N_h= Número de elementos en el estrato h.

N= Tamaño de la población.

$\bar{Y}_{k,h}$=Promedio de la variable transformada k en el estrato h

.

La varianza estimada de θ_k usando afijación proporcional al tamaño de muestra, es:
$$V\left(\hat{\theta_k}\right) = \frac{N-n}{N^2 n} \sum_{h=1}^{L} N_h S_{k,h}^2 \qquad k = 1, 2, \ldots, K.$$
la cual constituye una medida de calidad para la estratificación de la k-ésima variable.

Es posible demostrar que al aumentar el número de estratos, la varianza irá disminuyendo; con esto, se puede deducir que las estratifaciones con un número alto de estratos, darán mejores resultados que aquellas con un número bajo. Sin embargo, en la práctica se ha encontrado que para más de siete estratos la ganancia en la disminución de varianza no es muy significativa[1].

[1]En el caso de trabajar con una estratificación univariada, se sigue el procedimiento expuesto con K=1; se denota con $C_K^\star$ a la clasificación óptima de la variable k (obtenida por el método de Dalenius) y $V^\star(\theta_k)$ cálculada con la fórmula anterior.

Apéndice C

Definición de los indicadores.

A continuación se presenta la definición de cada indicador. A partir de los objetivos de las Verificaciones, se establecieron los indicadores que se obtendrían y se diseñaron los cuestionarios. Las fórmulas para el cálculo de éstos se presentan haciendo uso de la numeración de las preguntas de cada cuestionario y el código de la respuesta.

El cálculo de los indicadores se hace sobre los registros que presentan la características de interés o se conoce su situación específica; esto es, se excluyen los casos donde el informante no pudo responder: la no respuesta y el "*no sabe*"[1] .

Empadronados
Definición: Porcentaje de empadronados, respecto a la población de 18 años y más.
Fórmula: $\frac{9.1=1+11.1=1+(12.1=1 \text{y} 12.4=1,2ó3)}{\text{Población}-12.1=3-(12.1=1 \text{y} 12.4=4ó5)}$

[1] IFE, RFE, "Verificación Nacional Muestral 2005", agosto de 2005; Informe Final de Resultados, y "Verificación Nacional Muestral 2006", mayo de 2006; Informe Final de Resultados.

Empadronados en el estado

Definición: Porcentaje de empadronados en el estado, respecto a la población de 18 años y más.

Fórmula:

$$\frac{9.1{=}1{+}11.2.1{=}\textit{“Mismo Estado”}{+}12.2{=}2ó3{+}(12.2{=}1y12.4{=}2ó3){+}(12.2{=}1y12.5{=}1)}{\text{Población}{-}11.2{=}3{-}11.2.1{=}0{-}12.1{=}3{-}12.2{=}5{-}(12.2{=}1y12.4{=}4){-}(12.2{=}1y12.5{=}3)}$$

Empadronados en la sección

Definición: Porcentaje de empadronados en la sección electoral en la que residen, respecto a la población de 18 años y más.

Fórmula:

$$\frac{9.1{=}1{+}(11.2.1{=}\textit{“Mismo Estado”}\text{y}11.2.2{=}\textit{“Misma Sección”}){+}(12.2{=}1y{+}12.4{=}2ó3){+}(12.2{=}1y12.5{=}1)}{\text{Población}{-}(12.2{=}1y12.4{=}5){-}(12.2{=}1y12.5{=}3)}$$

Credencializados

Definición: Porcentaje de población empadronada que tiene Credencial para Votar, respecto a la población de 18 años y más.

Fórmula:

$$\frac{10.1{=}1{+}11.1{=}1}{\text{Población}{-}10.1{=}3{-}(11.1{=}3y12.1{=}2ó3)}$$

Credencializados en el estado

Definición: Porcentaje de población empadronada que tiene Credencial para Votar del estado donde reside, respecto a la población de 18 años y más.

Fórmula:

$$\frac{10.1{=}1{+}11.2.1{=}\textit{“Mismo Estado”}}{\text{Población}{-}10.1{=}3{-}11.2.1{=}0{-}(11.1{=}3y12.1{=}2){-}(11.1{=}3y12.2{=}1{,}2{,}3ó5)}$$

Credencializados en la sección

Definición: Porcentaje de población empadronada que tiene Credencial para Votar de la sección electoral donde reside, respecto a la población de 18 años y más.

Fórmula:

$$\frac{10.1{=}1{+}(11.2.1{=}\textit{“Mismo Estado”}\text{y}11.2.2{=}\textit{“Misma Sección”})}{\text{Población}{-}10.1{=}3{-}(11.1{=}3y12.2{=}1ó5)}$$

Fallecidos en el padrón

Definición: Porcentaje de registros en el padrón correspondientes a ciudadanos fallecidos, respecto al Padrón Electoral.

Fórmula: $\frac{5.1=3}{\text{Padrón}-1.1=4-2.1=10-3.1=2-5.1=4}$

Cambio de domicilio no reportado según sea el destino

Definición: Porcentaje de empadronados que cambiaron de domicilio y no lo han reportado, respecto al Padrón Electoral.

Fórmula: $\frac{5.1=1}{\text{Padrón}-1.1=4-2.1=10-3.1=2-5.1=4}$

Cambio de domicilio al mismo municipio

Definición: Porcentaje de empadronados que cambiaron de domicilio dentro del municipio de su registro y no lo han reportado, respecto al Padrón Electoral.

Fórmula: $\frac{5.1=1 \text{y} 6,2=1}{\text{Padrón}-1.1=4-2.1=10-3.1=2-5.1=4}$

Cambio de domicilio a otro municipio dentro del mismo estado

Definición: Porcentaje de empadronados que cambiaron de domicilio a otro municipio dentro del mismo estado de su registro y no lo han reportado, respecto al Padrón Electoral.

Fórmula: $\frac{5.1=1 \text{y} 6.2=2}{\text{Padrón}-1.1=4-2.1=10-3.1=2-5.1=4}$

Cambio de domicilio a otro estado

Definición: Porcentaje de empadronados que cambiaron de domicilio a un estado distinto al de su registro y no lo han reportado, respecto al Padrón Electoral.

Fórmula: $\frac{5.1=1 \text{y} 6.2=3}{\text{Padrón}-1.1=4-2.1=10-3.1=2-5.1=4}$

Cambio de domicilio a otro país

Definición: Porcentaje de empadronados que cambiaron de domicilio a otro país y no lo han reportado, respecto al Padrón Electoral.

Fórmula: $\frac{5.1=1 \text{y} 6.2=4}{\text{Padrón}-1.1=4-2.1=10-3.1=2-5.1=4}$

Apéndice D

Cuestionarios.

A continuación se anexa una copia de los dos cuestionarios utilizados en la VNM2006. Para la Encuesta de Cobertura, tiene el nombre de **"CUESTIONARIO DE SITUACIÓN DE CIUDADANOS EN EL PADRÓN"**; mientras que para la Encuesta de Actualización lleva el nombre de **"CÉDULA DE IDENTIFICACIÓN DE CIUDADANOS EN EL PADRÓN"**.

INSTITUTO FEDERAL ELECTORAL
REGISTRO FEDERAL DE ELECTORES

VERIFICACIÓN NACIONAL MUESTRAL, 2006
CUESTIONARIO DE SITUACIÓN DE CIUDADANOS EN EL PADRÓN

1. CARACTERÍSTICA DE LA LOCALIDAD POR LA QUE NO SE REALIZÓ LA ENTREVISTA

1. Deshabitada
2. Inexistente
3. Difícil acceso
4. Otra _____ (Especifica)

Código

Fin de llenado

HOJA ____ **DE** ____

FOLIO DE CAPTURA

NÚMERO DE TABLA

NÚMERO CONSECUTIVO DE VIVIENDA SELECCIONADA

2. IDENTIFICACIÓN GEOELECTORAL

ESTADO	DISTRITO	MUNICIPIO	SECCIÓN	LOCALIDAD	MANZANA

3. DOMICILIO DE LA VIVIENDA

CALLE _____ NUM. EXT _____ NUM. INT _____

COLONIA O LOCALIDAD _____

4. CARACTERÍSTICAS DEL PREDIO

Anota el código que corresponda

SÓLO APLICA PARA ZONA URBANA

1. Vivienda habitada (***pasa a 5***)
2. Hotel, pensión, campamento de trabajo, etc.
3. Vivienda deshabitada
4. Edificación en construcción
5. Uso distinto al de vivienda
6. Predio baldío
7. No se localizó

Fin de entrevista

Código

5. REALIZACIÓN DE LA ENTREVISTA

Anota el código que corresponda, según la situación de la entrevista

VISITAS

1ra.	2da	3ra.
Código	*Código*	*Código*

1. Sí se realizó la entrevista (***pasa a 6.1***)
2. No, por ausencia
3. No, por informante inadecuado
4. No, por rechazo

Programa visita

2da _____

3ra _____

6. OCUPANTES DE LA VIVIENDA

6.1 ¿Cuántas personas de 18 años de edad y mayores de 18 viven en esta vivienda?	6.2 ¿Cuántas personas cumplirán 18 años antes del 3 de julio de 2006?	6.3 Total (suma 6.1 más 6.2)
NÚMERO	NÚMERO	NÚMERO
Pasa a 6.2	***Pasa a 6.3***	***Pasa a 7***

7. RESIDENTES HABITUALES	8. RESIDENCIA ANTERIOR	9. BÚSQUEDA EN PADRÓN	10. ESTÁ EN PADRÓN			11. NO ESTÁ EN PADRÓN		12. SOLICITUD DE CREDENCIAL						13. INFORMANTE
¿Me podría dar el nombre de cada uno de los ciudadanos, empezando por el suyo?	¿Hace un año, en marzo de 2005, vivía en este domicilio?	Revisa si está en el Listado de Padrón	10.1 ¿Tiene su credencial para votar?	10.2 ¿Su credencial es de este domicilio?	10.3 ¿Su nombre y domicilio están correctos?	11.1 ¿Tiene su credencial para votar?	11.2 Solicítala y anota los siguientes datos	12.1 ¿Alguna vez ha solicitado su credencial para votar?	12.2 La última vez que solicitó su credencial dio... (***lee las opciones***)	12.3 ¿En qué estado la solicitó?	12.4 ¿Por qué causa no la tiene?	12.5 ¿Ha realizado algún trámite para solicitar otra credencial?	12.6 ¿Por qué motivo no lo ha realizado?	¿Quién informó?
1. _____ NOMBRE(S) _____ APELLIDO PATERNO APELLIDO MATERNO FECHA DE NAC (Día Mes Año) SEXO (H o M)		***Circula una opción*** 1. Sí está, ***anota consecutivo*** ____ ***Pasa a 10.1*** 2. No está ***(pasa a 11.1)***					ESTADO |__|__| SECCIÓN |__|__|__|__| ***No se obtuvo Información*** |__| Anota código 3			ESTADO |__|__|				
2. _____ NOMBRE(S) _____ APELLIDO PATERNO APELLIDO MATERNO FECHA DE NAC (Día Mes Año) SEXO (H o M)		***Circula una opción*** 1. Sí está, ***anota consecutivo*** ____ ***Pasa a 10.1*** 2. No está ***(pasa a 11.1)***					ESTADO |__|__| SECCIÓN |__|__|__|__| ***No se obtuvo Información*** |__| Anota código 3			ESTADO |__|__|				
3. _____ NOMBRE(S) _____ APELLIDO PATERNO APELLIDO MATERNO FECHA DE NAC (Día Mes Año) SEXO (H o M)		***Circula una opción*** 1. Sí está, ***anota consecutivo*** ____ ***Pasa a 10.1*** 2. No está ***(pasa a 11.1)***					ESTADO |__|__| SECCIÓN |__|__|__|__| ***No se obtuvo Información*** |__| Anota código 3			ESTADO |__|__|				
OBSERVACIONES	1. Sí 2. No 3. No sabe 4. No contestó ***Pasa a 9***		1. Sí ***(pasa a 10.2)*** 2. No ***(pasa a 12.4)*** 3. No sabe 4. No contestó ***Pasa a 13***	1. Sí ***(pasa a 10.3)*** 2. No ***(pasa a 12.5)*** 3. No sabe 4. No contestó ***Pasa a 13***	1. Sí ***(pasa a 13)*** 2. No ***(pasa a 12.5)*** 3. No sabe 4. No contestó ***Pasa a 13***	1. Sí ***(pasa a 11.2)*** 2. No 3. No sabe 4. No contestó ***Pasa a 12.1***	***Pasa a 12.5***	1.Sí ***(pasa a 12.2)*** 2. No ***(pasa a 12.6)*** 3. No sabe 4. No contestó ***Pasa a 13***	1. Este domicilio 2. Otro domicilio del municipio 3. Otro municipio del estado 4. Otro estado ***(pasa a 12.3)*** 5. No sabe 6. No contestó ***Pasa a 12.4***	Consulta tabla de estados ***Pasa a 12.4***	1. Extravío o robo ***(pasa a 12.5)*** 2. No está en módulo 3. No la ha recogido 4. Otra causa 5. No sabe 6. No contestó ***Pasa a 13***	1. Sí ***(pasa a 13)*** 2. No ***(pasa a 12.6)*** 3. No sabe 4. No contestó ***Pasa a 13***	1. No tiene tiempo 2. No le interesa 3. No desea actualizar su domicilio 4. No sabe cómo solicitarla 5. Desconoce dónde está el módulo 6. El servicio de módulo es lento 7. Es extranjero 8. Otra causa 9. No sabe ***Pasa a 13***	1. Ciudadano en cuestión Informante familiar ***(Anota parentesco)*** 3. Informante no familiar

7. RESIDENTES HABITUALES	8. RESIDENCIA ANTERIOR	9. BÚSQUEDA EN PADRÓN	10. ESTÁ EN PADRÓN			11. NO ESTÁ EN PADRÓN		12. SOLICITUD DE CREDENCIAL						13. INFORMANTE
¿Me podría dar el nombre de cada uno de los ciudadanos, empezando por el suyo?	¿Hace un año, en marzo de 2005, vivía en este domicilio?	Revisa si está en el Listado de Padrón	10.1 ¿Tiene su credencial para votar?	10.2 ¿Su credencial es de este domicilio?	10.3 ¿Su nombre y domicilio están correctos?	11.1 ¿Tiene su credencial para votar?	11.2 Solicítala y anota los siguientes datos	12.1 ¿Alguna vez ha solicitado su credencial para votar?	12.2 La última vez que solicitó su credencial dio... (***lee las opciones***)	12.3 ¿En qué estado la solicitó?	12.4 ¿Por qué causa no la tiene?	12.5 ¿Ha realizado algún trámite para solicitar otra credencial?	12.6 ¿Por qué motivo no lo ha realizado?	¿Quién informó?
4. NOMBRE(S) / APELLIDO PATERNO / APELLIDO MATERNO / FECHA DE NAC (Día Mes Año) / SEXO (H o M)		***Circula una opción*** 1. Sí está, ***anota consecutivo*** _____ ***Pasa a 10.1*** 2. No está ***(pasa a 11.1)***					____ ESTADO ____ SECCIÓN ***No se obtuvo Información*** ____ Anota código 3			____ ESTADO				
5. NOMBRE(S) / APELLIDO PATERNO / APELLIDO MATERNO / FECHA DE NAC (Día Mes Año) / SEXO (H o M)		***Circula una opción*** 1. Sí está, ***anota consecutivo*** _____ ***Pasa a 10.1*** 2. No está ***(pasa a 11.1)***					____ ESTADO ____ SECCIÓN ***No se obtuvo Información*** ____ Anota código 3			____ ESTADO				
6. NOMBRE(S) / APELLIDO PATERNO / APELLIDO MATERNO / FECHA DE NAC (Día Mes Año) / SEXO (H o M)		***Circula una opción*** 1. Sí está, ***anota consecutivo*** _____ ***Pasa a 10.1*** 2. No está ***(pasa a 11.1)***					____ ESTADO ____ SECCIÓN ***No se obtuvo Información*** ____ Anota código 3			____ ESTADO				
7. NOMBRE(S) / APELLIDO PATERNO / APELLIDO MATERNO / FECHA DE NAC (Día Mes Año) / SEXO (H o M)		***Circula una opción*** 1. Sí está, ***anota consecutivo*** _____ ***Pasa a 10.1*** 2. No está ***(pasa a 11.1)***					____ ESTADO ____ SECCIÓN ***No se obtuvo Información*** ____ Anota código 3			____ ESTADO				
8. NOMBRE(S) / APELLIDO PATERNO / APELLIDO MATERNO / FECHA DE NAC (Día Mes Año) / SEXO (H o M)		***Circula una opción*** 1. Sí está, ***anota consecutivo*** _____ ***Pasa a 10.1*** 2. No está ***(pasa a 11.1)***					____ ESTADO ____ SECCIÓN ***No se obtuvo Información*** ____ Anota código 3			____ ESTADO				
OBSERVACIONES	1. Sí 2. No 3. No sabe 4. No contestó ***Pasa a 9***		1. Sí (***pasa a 10.2***) 2. No (***pasa a 12.4***) 3. No sabe → ***Pasa a 13*** 4. No contestó → ***Pasa a 13***	1. Sí ***(pasa a 10.3)*** 2. No ***(pasa a 12.5)*** 3. No sabe → ***Pasa a 13*** 4. No contestó → ***Pasa a 13***	1. Sí ***(pasa a 13)*** 2. No ***(pasa a 12.5)*** 3. No sabe → ***Pasa a 13*** 4. No contestó → ***Pasa a 13***	1. Sí ***(pasa a 11.2)*** 2. No 3. No sabe → ***Pasa a 12.1*** 4. No contestó → ***Pasa a 12.1***	***Pasa a 12.5***	1.Sí ***(pasa a 12.2)*** 2. No ***(pasa a 12.6)*** 3. No sabe → ***Pasa a 13*** 4. No contestó → ***Pasa a 13***	1. Este domicilio 2. Otro domicilio del municipio 3. Otro municipio del estado → ***Pasa a 12.4*** 4. Otro estado ***(pasa a 12.3)*** 5. No sabe → ***Pasa a 12.4*** 6. No contestó → ***Pasa a 12.4***	Consulta tabla de estados ***Pasa a 12.4***	1. Extravío o robo ***(pasa a 12.5)*** 2. No está en módulo 3. No la ha recogido 4. Otra causa 5. No sabe 6. No contestó → ***Pasa a 13***	1. Sí ***(pasa a 13)*** 2. No ***(pasa a 12.6)*** 3. No sabe → ***Pasa a 13*** 4. No contestó → ***Pasa a 13***	1. No tiene tiempo 2. No le interesa 3. No desea actualizar su domicilio 4. No sabe cómo solicitarla 5. Desconoce dónde está el módulo 6. El servicio de módulo es lento 7. Es extranjero 8. Otra causa 9. No sabe ***Pasa a 13***	1. Ciudadano en cuestión Informante familiar ***(Anota parentesco)*** 3. Informante no familiar

NOMBRE, FIRMA Y CLAVE DEL VISITADOR DOMICILIARIO (P.P.)			NOMBRE, FIRMA Y CLAVE DEL VISITADOR DOMICILIARIO (R.F.E.)		
NOMBRE	**FIRMA**	**CLAVE**	**NOMBRE**	**FIRMA**	**CLAVE**

INSTITUTO FEDERAL ELECTORAL
REGISTRO FEDERAL DE ELECTORES

VERIFICACIÓN NACIONAL MUESTRAL, 2006
CÉDULA DE IDENTIFICACIÓN DE CIUDADANOS EN EL PADRÓN

CONSECUTIVO	***01242***
ENTIDAD	***NUEVO LEÒN 19***
DISTRITO	***05***
MUNICIPIO	***MONTERREY 040***
LOCALIDAD	***MONTERREY 0001***
SECCION	***1470***
MANZANA	***7***
NOMBRE	***RAMIREZ CALZADA GABRIELA***
DOMICILIO	***AV. EUGENIO GARZA SADA # 2125***
	COL. ROMA

EDAD ACTUAL: ***24*** *FECHA DE TRÁMITE*: ***19 / 02 / 2003***

Si es duplicado, anota el consecutivo de la cédula con información _____ *CONSECUTIVO*

ENCIERRA EN UN CÍRCULO LA OPCIÓN CORRESPONDIENTE Y ANOTA EL CÓDIGO EN EL RECUADRO

1. LOCALIZACIÓN DEL DOMICILIO

1.1 El domicilio del ciudadano:

1. Es el mismo del padrón
2. Tiene algún error, pero pertenece a la sección
3. Está en otra sección _____ *CLAVE*
} ***Pasa a 2***

4. No se localizó el domicilio ***(pasa a 1.2)***

Código

1.2 Pregunta a los vecinos si reconocen al ciudadano en cuestión

Sí lo reconocen y saben en donde vive ***(regresa a 1.1)***

1. Sí lo reconocen, pero no saben en dónde vive ***(pasa a 7.1)***
2. No lo reconocen ***(pasa a 3.2)***

Código

2. CONDICIÓN DE LA ENTREVISTA

Anota el código que corresponda, según la situación de la entrevista

V I S I T A S

1ra	2da		3ra	
Código	Código	Fecha _____	Código	Fecha _____
		Hora _____		Hora _____

1. Entrevista con el ciudadano en cuestión ***(pasa a 4)***

2. El ciudadano en cuestión no se encontró al momento de la entrevista — ***Programa cita; en 3ra visita realiza la entrevista con un informante del domicilio y pasa a 3.1***

3. No había nadie en la vivienda
4. Informante inadecuado
5. Rechazo
} ***Programa cita; en 3ra visita habilita a un vecino como informante y pasa a 3.1***

6. Vivienda deshabitada
7. Vivienda de uso temporal
8. Uso distinto al de vivienda
9. Predio baldío
} ***Habilita a un vecino como informante y pasa a 3.1***

10. No hubo informante ***(Fin de entrevista)***

3. RECONOCIMIENTO DEL CIUDADANO EN CUESTIÓN

3.1 Pregunta si reconocen al ciudadano en cuestión

1. Sí lo reconocen ***(pasa a 4)***
2. No lo reconocen ***(pasa a 3.2)***

Código

3.2 Anota cuántos vecinos consultaste y en la parte de atrás de la cédula registra el nombre y domicilio de cada uno.

_____ *NÚMERO*

Pasa a 7.1

4. LUGAR DE RESIDENCIA DEL CIUDADANO EN CUESTIÓN

¿Vive *(menciona el nombre del ciudadano)* en *(menciona el domicilio localizado en campo)*?

1. Sí ***(pasa a 7.1)***
2. No ***(pasa a 5)***

Código

5. CAUSA DE NO RESIDENCIA

¿Por qué no vive aquí el ciudadano?

1. Cambio de domicilio ***(pasa a 6.1)***
2. Nunca ha vivido aquí ***(pasa a 6.2)***
3. Falleció
4. No sabe
} ***Pasa a 7.1***

Código

6. CAMBIO DE DOMICILIO DEL CIUDADANO EN CUESTIÓN

6.1 ¿Desde cuándo no vive aquí el ciudadano?

1. _____ *AÑOS* **2.** Menos de un año **3.** No sabe

Pasa a 6.2

Código

6.2 ¿El ciudadano vive... *(lee las opciones)*?

1. Dentro del municipio ***(pasa a 6.3)***
2. En otro municipio, dentro del estado
3. En otro estado
} ***Pasa a 7.1***

4. En otro país ***(pasa a 6.4)***
5. No sabe ***(pasa a 7.1)***

Código

6.3 ¿Me podría dar el domicilio actual del ciudadano?

1. Sí _____ *CALLE*

_____ *NUM. EXT* _____ *NUM. INT* _____ *COLONIA O LOCALIDAD*

2. Se negó a proporcionar la información
3. No sabe ***(Consulta a otros vecinos)*** _____ *NÚMERO*

Pasa a 7.1

Código

6.4 ¿En qué país y ciudad vive?

1. Estados Unidos _____ *CIUDAD*
2. Canadá _____ *CIUDAD*
3. Otro país _____ *PAÍS* _____ *CIUDAD*
4. No sabe

Pasa a 7.1

Código

7. DATOS DEL INFORMANTE

7.1 ¿Cuánto tiempo tiene de vivir en el domicilio?

1. _____ *AÑOS* _____ *MESES* ***(pasa a 7.2)***
2. Es el ciudadano en cuestión y no vive en el domicilio ***(pasa a 7.4)***

Código

7.2 El informante es...

1. El ciudadano en cuestión ***(pasa a 7.4)***
2. Padre o Madre del ciudadano en cuestión
3. Esposo(a) del ciudadano en cuestión
4. Hermano(a) del ciudadano en cuestión
5. Hijo(a) del ciudadano en cuestión
6. Abuelo(a) del ciudadano en cuestión
7. Otro parentesco _____ *ESPECIFICA*
8. Informante no familiar ***(si la entrevista la realizas con un vecino anota su domicilio)***
} ***Pasa a 7.3***

Código

_____ *CALLE* _____ *NUM. EXT.* _____ *NUM. INT.*

7.3 Nombre completo del informante

_____

Pasa a 7.4

7.4 Solicita la firma o huella al informante

8. PARA USO DEL TÉCNICO CARTÓGRAFO

Revisa el domicilio de la preg. 6.3

1. El domicilio está dentro de la sección
2. El domicilio está fuera de la sección
3. No se pudo definir

Código

INSTITUTO FEDERAL ELECTORAL
REGISTRO FEDERAL DE ELECTORES

CIUDADANOS QUE VIVEN EN EL DOMICILIO LOCALIZADO PERO NO PROPORCIONARON SU INFORMACION PERSONALMENTE

1. ¿Por qué motivo no se encuentra el ciudadano en la vivienda?

1. Trabaja
2. Estudia
3. Está enfermo u hospitalizado
4. Otra causa ---------- *ESPECIFICA*

Código

Pasa a 2

2. ¿En dónde se encuentra el ciudadano?

1. Está en la ciudad o municipio
2. Está en el país, pero fuera de la ciudad o municipio } ***Pasa a 4***
3. Está fuera del país *(pasa a 3)*
4. No sabe (pasa a 4)

Código

3. ¿En qué país y ciudad?

1. Estados Unidos ---------- *CIUDAD*
2. Canadá ---------- *CIUDAD*
3. Otro país ---------- *PAÍS* , ---------- *CIUDAD*
4. No sabe

Código

Pasa a 4

4. ¿Desde cuándo está ausente el ciudadano?

1. Menos de un mes
2. Menos de un año ___ *MESES*
3. Un año o más ___ *AÑOS*
4. No sabe

Código

Pasa a 5

5. ¿El ciudadano viene a este domicilio cada...? (Lee las opciones)

1. Día
2. Semana
3. Mes } ¿Cuántas veces? ___ *NÚMERO*
4. Año ¿En qué periodo?

Circula los meses que te indiquen

1. Ene **2**. Feb **3**. Mar **4**. Abr **5**. May **6**. Jun **7**. Jul. **8**. Ago **9**. Sep **10**. Oct **11**. Nov **12**. Dic

5. Hace más de un año que no viene ___ *AÑOS*
6. No sabe

Código

Fin de la entrevista

RELACIÓN DE VECINOS CONSULTADOS (VIENE DE LA PREGUNTA 3.2)

NOMBRE	DOMICILIO
NOMBRE	DOMICILIO
NOMBRE	DOMICILIO

OBSERVACIONES

NOMBRE, FIRMA Y CLAVE DEL VISITADOR DOMICILIARIO

NOMBRE	FIRMA	CLAVE

Bibliografía

[1] Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Jhon Wiley & Sons.

[2] Duran, B. and Odell, P. (1974). *Cluster Analysis*. Springer Verlag, Berlin.

[3] Graybill, F. (1976). *Theory and Application of the Linear Model*. Duxbury Classic.

[4] Graybill, F. (1976). *An Introduction to Linear Statistical Models*. McGraw-Hill.

[5] Hartigan, J. A. (1967). *Distribution of the residual sum of squares in fitting inequalities*. Biometrika. 54, pp. 69 - 84.

[6] Hartigan, J. A. (1967). *Representation of similarity matrices by trees*. Journal of the American Statistical Association. 62, pp. 1140-1158.

[7] Hartigan, J. A. (1975). *Clustering Algorithms*. John Wiley & Sons, Inc. New York, NY, US.

[8] Huberty, C. T. (1994). *Applied Discriminant Analysis*. John Wiley & Sons. New York, NY, US.

[9] Jolliffe, I. T. (1986). *Principal Component Analysis*. Second Edition. Springer Verlag, Inc. New York, NY, US.

[10] Jolliffe, I. T. (1987a). *Selection of variables*. Applied Statistics. 36, pp. 373-374.

[11] Mardia, K. V., Kent, J. T., and Bybbi, J. M. (1979). *Multivariate Analysis*. Academic Press: London.

[12] Mood, A., Graybill, F. and Boes, D. (1974). *Introduction to the Theory of Statistics*. McGraw-Hill Series in Probability and Statistics.

[13] Morrison, D. (1990). *Multivariate Statistical Methods*. McGraw-Hill.

[14] Peña, D. (2002). *Análisis de Datos Multivariantes*. Editorial McGraw-Hill.

[15] Seber, G. A. F. (1984). *Multivariate Observations*. Wiley series in probability and mathematical statistics. John Wiley & Sons.

[16] Tsay, R. S. (1988). *Outliers, level shifts, and variance changes in time series*. Journal of Forecasting, Vol. 7, pp. 1-20.

[17] Ward, J. H. (1963). *Hierarchical grouping to optimizer an objective function*. Journal of the American Statistical Association. 53, pp. 236-244.

[18] Cartografía proporcionada por el Instituto Nacional de Estadística Geográfica e Informática (INEGI).