00365

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

POSGRADO EN CIENCIAS MATEMATICAS

FACULTAD DE CIENCIAS

Técnicas estadísticas de solución al problema de agrupación. Aplicación a la clasificación de la población por su condición de pobreza.

T E S I S

QUE PARA OBTENER EL GRADO ACADEMICO DE

MAESTRO EN CIENCIAS MATEMATICAS

P R E S E N T A

HUMBERTO SOTO DE LA ROSA

DIRECTOR DE TESIS:

DR. RAUL RUEDA DIAZ DEL CAMPO

MEXICO, D. F. SEPTIEMBRE, 2004

Índice:

Agradecimientos

A mi amada esposa, por su apoyo, comprensión e infinito amor.

A mi familia, que ha estado a mi lado siempre.

A mis amigos, compañeros fieles en el sendero de la vida.

A mi asesor Raúl Rueda por su tiempo, al igual que a los sinodales y demás maestros que contribuyeron a que este trabajo fuera posible.

Principalmente gracias al ser supremo, por permitirme llegar hasta este punto del camino de la vida.

Motivación del Proyecto y Objetivos

a. Motivación del proyecto

Este proyecto de tesis está basado en una visión sobre la situación económica que se presenta a finales del siglo XX y principios del XXI en los países en vías de desarrollo y en particular en el caso de México. La visión referida plantea una serie de problemáticas que surgen de procesos como la globalización, que obligan a los países a alinearse a esquemas de desarrollo que si bien pueden ser benéficos para las economías en su conjunto, llegan a "castigar" de manera significativa a ciertos sectores de la población, que son en ocasiones los últimos beneficiados de los esquemas de desarrollo.

La población de los sectores que resienten la problemática planteada es generalmente aquélla con menor capacidad de respuesta a procesos inflacionarios o al aumento en los niveles de desempleo por ejemplo, lo que la convierte en vulnerable hacia muchos aspectos negativos de la vida cotidiana. Una de las propuestas que se han planteado para contrarrestar los efectos negativos de las políticas de desarrollo económico en la población vulnerable a ellos es la implementación de acciones de "política social", que consisten en la asignación de recursos destinados a atenuar los problemas que afectan a la población referida[1] (Gallardo y Osorio, 1998).

Para implementar la "política social", es necesario identificar primero quiénes son los individuos que conforman a la población vulnerable, a quienes irán dirigidas sus acciones. Esto plantea la necesidad específica de generar una agrupación de la población de acuerdo a su condición de vulnerabilidad y por ende de pobreza como su más clara manifestación. Dicha agrupación requiere de gran precisión pues errores cometidos en ella pueden provocar que los recursos no sean utilizados de manera óptima.

b. El problema de clasificación de la población por su condición de pobreza

Debido a lo improbable de tener un censo cuando se busca generar una agrupación de la población por su condición de pobreza (por motivo de costos, distancias, tiempos,

[1] La "política social" puede ser aplicada a la población en general, de hecho existe una amplia discusión entre las ventajas y desventajas de implementar acciones de manera generalizada con respecto a aplicarla de manera focalizada.

entre otras limitantes), se dispondrá generalmente de una muestra. En muchas ocasiones es conveniente en la práctica que a partir de la información proveniente de la muestra se genere una regla de clasificación que permita identificar el grupo al que pertenece un individuo que no estaba considerado en la muestra original. Esto se lleva a cabo, bajo el supuesto de que se tiene una muestra aleatoria de la población, utilizando técnicas de clasificación predictivas que permitan generar una regla para poder agrupar a los individuos nuevos. Algunas técnicas que se han propuesto desde el punto de vista estadístico son el Análisis Discriminante, la Regresión Logística y algunas propuestas con enfoque Bayesiano.

Las técnicas de clasificación predictivas parten del supuesto de que se tiene una agrupación inicial conocida, la cual se intenta replicar a partir de un conjunto de variables que contienen información relevante relacionada con la agrupación. De hecho a partir de la generación de la regla de clasificación se pueden conocer cuáles variables tienen mayor nivel de injerencia (estadísticamente significativa) en la agrupación.

La necesidad de conocer una agrupación inicial conduce a tener una definición de los grupos, en este caso por su condición de pobreza, reiterando la importancia de que esta definición sea lo más precisa posible. Por ello es indispensable estudiar las propuestas de agrupación inicial de que se dispone y seleccionar una de ellas.

c. El problema de agrupación

Obtener una agrupación de la población de acuerdo con su condición de pobreza plantea la necesidad de definir el concepto de pobreza, de establecer la unidad de análisis, de ubicar la fuente de información o en su caso de generarla, de identificar los indicadores para medir dicho concepto y de aplicar la metodología seleccionada. Existe un gran número de combinaciones posibles, derivadas tanto de las opciones a elegir como de la subjetividad que puede implicarse, lo cual conduce a que las agrupaciones obtenidas tengan diferencias entre sí, provocando que el problema sea sumamente complejo (Comité Técnico para la Medición de la Pobreza, 2002).

Existen muy variadas propuestas metodológicas, planteadas a partir de una visión económica, que dan respuesta a la necesidad de obtener una agrupación de la población de acuerdo con su condición de pobreza, tema que ha sido extensamente estudiado desde el punto de vista económico en los años recientes. Propuestas que han variado a través del tiempo, y entre las que se encuentran, solo por mencionar las más renombradas, la visión del ingreso o gasto, o la de los funcionamientos y las capacidades o potencialidades (Sen, 1997).

El problema puede ser estudiado también desde un punto de vista estadístico por medio de técnicas de agrupación como el análisis de conglomerados jerárquicos o el análisis de conglomerados de medias móviles, así como algunas otras alternativas que con poca frecuencia se utilizan, como la técnica denominada "Simulated Annealing".

d. Objetivos del proyecto

Antes de plantear los objetivos del presente trabajo es necesario señalar que en el año 2002 el gobierno de México convocó a un grupo de expertos en la materia de análisis y medición de la pobreza, denominado Comité Técnico para la Medición de la Pobreza, para que juntos definieran la mejor estrategia posible para obtener una agrupación de los hogares por su condición de pobreza. Los resultados obtenidos condujeron a los analistas a definir la pobreza como una privación de los elementos necesarios para la vida humana dentro de una sociedad, así como de medios o recursos para modificar esta situación. Los miembros del Comité decidieron utilizar una metodología basada en la medición del ingreso y comparación del mismo con una línea de pobreza a nivel de hogar dado que la encuesta disponible para tal efecto, la Encuesta Nacional de Ingreso y Gasto de los Hogares (ENIGH), tiene justamente esa unidad de análisis.

Como se ha señalado, los problemas de subjetividad en las definiciones, así como de diversidad en las opciones de selección de metodologías provocan discrepancias entre resultados, lo cual es reconocido por el Comité así como lo es la posibilidad y necesidad de explorar alternativas, y la utilización reducida de técnicas estadísticas para buscar soluciones al problema de clasificación de la población de acuerdo con su condición de pobreza.

La pobreza es actualmente un concepto reconocido como multidimensional (López-Calva y Rodríguez Chamussy, 2004), y al analizarlo solo mediante un indicador pierde ese carácter. Aunado a esto, los riesgos de incorrecta captación del indicador, así como la sensitividad que puede tener la medición por ingreso con respecto a la pobreza de corto plazo al no considerar características estructurales, conduce a plantear la necesidad de evaluar los resultados del Comité, y en su caso, proponer alternativas más eficientes.

El presente proyecto busca responder a los siguientes tres objetivos:

1) evaluar la propuesta de agupación de la población de acuerdo con su condición de pobreza elaborada por el Comité, empleando para ello tres técnicas estadísticas aplicadas

a resolver el problema de clasificación de la población de acuerdo con su condición de pobreza, el Análisis Discriminante, la Regresión Logística Multinomial y una propuesta dentro del enfoque Bayesiano;

2) proponer, si es posible, una agrupación de la población de acuerdo con su condición de pobreza, más eficiente a la planteada por el Comité, para lo cual se utilizará la técnica denominada "Simulated Annealing", y;

3) proponer una regla de clasificación para la población en México de acuerdo a la condición de pobreza, utilizando para ello la mejor alternativa de agrupación inicial de que se disponga de acuerdo con los resultados del primer y segundo objetivo, la cual puede ser la del Comité o la que resulte de la aplicación del "Simulated Annealing", y utilizando también la mejor alternativa de técnica de clasificación de entre las propuestas en el planteamiento del primer objetivo, esto es, el Análisis Discriminante, la Regresión Logística Multinomial o la propuesta dentro del enfoque Bayesiano.

e. Estructura del proyecto

Para responder a los objetivos trazados el presente trabajo se estructura en cuatro capítulos de la siguiente forma:

1) el primer capítulo contempla la identificación de un conjunto de datos confiable para la realización del trabajo, realizando el análisis exploratorio de las variables que reflejen el perfil de la población;

2) el segundo capítulo mostrará las propuestas históricas para la solución al problema de clasificación de los individuos por su condición de pobreza, junto con los resultados de algunas de ellas al aplicarlas al conjunto de datos seleccionados;

3) el tercer capítulo muestra las técnicas estadísticas propuestas para la solución al problema de clasificación, así como otras técnicas que serán aplicadas para resolver el referido problema, estableciendo su sustento formal;

4) el cuarto capítulo presenta los resultados de la aplicación de las técnicas para alcanzar los objetivos propuestos;

finalmente se incluye una sección de conclusiones, de referencias bibliográficas, y apéndices incluyendo tablas de resultados, demostraciones y programas.

Es importante señalar que, como se mostrará en el cuarto capítulo, se hará uso de datos en dos momentos de tiempo distintos con la finalidad de realizar una validación de la fiabilidad de los resultados obtenidos.

Para concluir esta sección introductoria, se debe mencionar que a excepción de los análisis obtenidos a partir de cálculos propios por medio de programación en el paquete MATLAB, correspondientes a la propuesta con enfoque Bayesiano y a la metodología de "Simulated Annealing", el resto de los análisis y cálculos estadísticos se realizó con el paquete SPSS.

Capítulo I. El perfil de la población en México

a. Fuentes de Datos

En el caso específico de México una encuesta que se tiene disponible, conteniendo información de diversas características a nivel nacional, es la Encuesta Nacional de Ingreso y Gasto de los Hogares (ENIGH[2]) del año 2000 y del año 2002. Esta encuesta incluye información sobre las características socioeconómicas de los hogares del país, sobre la condición de las viviendas y la posesión de enseres, así como sobre características de los integrantes de los hogares, y una sección muy completa y confiable de datos sobre el ingreso y el consumo de los hogares. Esta encuesta será considerada para la realización del presente proyecto por dos razones, en primer lugar por su nivel de confiabilidad y credibilidad, y en segundo lugar porque es la misma encuesta que fué utilizada por el Comité Técnico de Medición de la Pobreza para agrupar a los hogares por su condición de pobreza, y dado que un objetivo del presente trabajo es evaluar dicha agrupación, es pertinente hacerlo con los mismos datos. La idea es aplicar las técnicas estadísticas de análisis para la ENIGH 2000, obteniendo las reglas de clasificación de hogares por su condición de pobreza que se aplicarán a la ENIGH 2002 para verificar la fiabilidad de los resultados.

En relación con los aspectos metodológicos de las encuestas es conveniente señalar que de acuerdo con el INEGI, la Encuesta Nacional de Ingreso y Gasto de los Hogares está basada en la consideración de que el monto del ingreso, su procedencia y su forma de distribución condicionan en gran medida el nivel de bienestar de la población, puesto que es dicho ingreso el que determina la capacidad económica de los hogares para adquirir los bienes y servicios necesarios. Por ello, para abordar el estudio del monto, la procedencia y la distribución del ingreso y el gasto de los hogares, se seleccionó a la vivienda particular como unidad de muestreo, y a los hogares, sus miembros y la vivienda en sí como unidades de observación y análisis.

El marco conceptual de la ENIGH está basado en las recomendaciones internacionales de las Naciones Unidas y la Organización Internacional del Trabajo, y está articulado al Sistema de Cuentas Nacionales y a las encuestas de hogares que levanta el INEGI.

[2] El Instituto Nacional de Estadística, Geografía e Informática (INEGI) es el encargado de realizar el levantamiento de la ENIGH.

La ENIGH está diseñada para presentar información representativa tanto a nivel nacional como para dos estratos más, uno constituido por localidades de 2500 y más habitantes y otro con localidades de menos de 2500 habitantes.

La recolección de la información se llevó a cabo por medio de la visita a las viviendas seleccionadas, utilizando instrumentos de captación especializados, con un equipo de entrevistadores y supervisores capacitados de manera exhaustiva durante un mes sobre los procedimientos, lineamientos y criterios establecidos para la Encuesta.

Las unidades que conformaron la muestra de la ENIGH se seleccionaron con criterios probabilísticos con el propósito de asegurar que pudieran estimarse los indicadores correspondientes para todos los hogares. El esquema de muestreo es estratificado y bietápico. En la primera etapa se eligieron grupos de viviendas y en la segunda se seleccionó directamente a la vivienda. Este esquema de muestreo permite estimar distribuciones porcentuales con precisión y confianza aceptables. La muestra elegida tiene un tamaño de 10,000 hogares.

b. Perfil de la población en México en el año 2000

Como primer paso para alcanzar los objetivos planteados en el presente trabajo es necesario conocer la información disponible y con ello diseñar de manera efectiva el esquema de trabajo para obtener el mayor aprovechamiento de la información referida. Cabe destacar que de entre los datos que contiene la encuesta, se reconocen indicadores directos y se tiene la posibilidad de construir indicadores indirectos para identificar a la población en condiciones de pobreza por medio de ellos.

Utilizando técnicas de estadística descriptiva se realiza un análisis exploratorio de las variables de la encuesta. En el caso de indicadores de tipo cualitativo[3], se utilizará como estadística la frecuencia de cada categoría del indicador. A continuación se presenta la tabla de frecuencias[4] para el indicador del material predominante en los muros de las viviendas, por ejemplo.

[3] Indicadores que muestran la pertenencia de cada unidad de análisis a un grupo definido a partir de cierta cualidad.

[4] La tabla de frecuencias muestra el número de casos que existen en una población para cada categoría de una variable de tipo categórico o cualitativo. También muestra el porcentaje, que se denomina frecuencia relativa.

Material de los muros				
	Frecuencia	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa acumulada
Cartón, hule, tela, llantas, etc.	9,695	0.01	9,695	0.01
Lámina de cartón	62,410	0.3	72,105	0.3
Carrizo, bambú, palma o tejamanil	167,927	0.7	240,032	1
Embarro o bejaraque	258,250	1.1	498,282	2.1
Lámina de asbesto	3,860	0	502,142	2.2
Lámina metálica, fibra de vidrio, plástico o mica	122,471	0.5	624,613	2.7
Madera	1,394,346	6	2,018,959	8.7
Vidrio o cristal	1,856	0	2,020,815	8.7
Panel de concreto	13,706	0.1	2,034,521	8.8
Concreto monolítico	69,577	0.3	2,104,098	9.1
Adobe	2,066,573	8.9	4,170,671	18
Tabique, ladrillo, tabicón o block	18,655,912	80.5	22,826,583	98.4
Piedra o cemento (incluye cantera)	317,440	1.4	23,144,023	99.8
Otros materiales	42,035	0.2	23,186,058	100
Total	23,186,058	100		

Esta tabla de frecuencias muestra que, si bien la gran mayoría de los hogares en el país tiene muros de tabique, ladrillo, tabicón o block (80.5 %), aún hay un pequeño pero no por ello menos importante porcentaje de hogares con muros que van desde el cartón y el carrizo hasta la lámina metálica y la madera. Se podría pensar que una mayor proporción de los hogares cuyas viviendas tienen muros de material endeble se encuentran en condiciones de pobreza, con respecto a la proporción de hogares cuyas viviendas tienen muros de material firme que se encuentra en dichas condiciones.

En lo que respecta a indicadores de tipo cuantitativo[5] se utilizarán las estadísticas descriptivas media y desviación estándar en su carácter de medidas de tendencia central y de dispersión respectivamente. Para ejemplificar el análisis exploratorio de este tipo de indicadores se utilizará el relativo a los ingresos per cápita dado que existe un consenso general entre los estudiosos del tema desde el punto de vista económico que señala al ingreso como uno de los indicadores que mejor refleja la calidad de vida de los hogares, y en este caso es de gran utilidad para definir aquéllos hogares en condiciones de pobreza. La media del ingreso mensual per cápita es de 1,967.4 pesos con una desviación estándar de 4,268.2. En este momento no se está en condición de determinar si el ingreso referido es alto o bajo, mas adelante se presentará una medida de comparación para poder concluir al

[5] Indicadores que reflejan un atributo medible numéricamente para cada unidad de análisis de una población.

respecto, sin embargo, la desviación estándar es relativamente alta (coeficiente de variación mayor a 2), lo cual da un indicio de una alta dispersión que se traduce en desigualdad en la distribución del ingreso referido.

Como se puede observar, al analizar las estadísticas descriptivas del resto de los indicadores se puede obtener un perfil general de la población en México a finales del siglo XX y principios del siglo XXI. En los cuadros siguientes se presentan las frecuencias relativas para los indicadores de tipo categórico, y para los indicadores de tipo continuo la media y la desviación estándar así como el histograma correspondiente.

El cuadro siguiente muestra que, si bien prácticamente la mitad de los hogares del país se encuentran ubicados en zonas urbanas de más de 100 mil habitantes, aún existe un 23 por ciento de los mismos ubicados en zonas rurales. Dichas zonas frecuentemente se encuentran localizadas en lugares alejados de las zonas urbanas por lo cual el acceso a servicios básicos, educativos, de salud, entre otros, es en algunas ocasiones complicado.

Características generales	
Tamaño de la localidad de residencia	
Loc. de 100,000 hab. y más	49.7%
Loc. de 15,000 a 99,000 hab.	13.9%
Loc. de 2,500 a 14,999 hab.	13.5%
Loc de menos de 2,500 hab.	23.0%

Con respecto a los servicios se puede apreciar que existe alrededor de una cuarta parte de los hogares del país que no tienen servicios como drenaje o recolección pública de basura, lo cual puede relacionarse con condiciones insalubres y por tanto propensión a enfermedades. En cuanto al servicio de energía eléctrica, prácticamente la totalidad del país tiene acceso al mismo, lo cual no ocurre con la disponibilidad de agua, para la cual aún existe un 9 por ciento sin dicho servicio, implicando la necesidad de adquirir el agua por medios costosos o, en el peor de los casos teniendo que acarrearla de distancias considerables.

Servicios		
Drenaje		73.3%
Recolección pública de basura		75.7%
Luz		98.0%
Disponibilidad de agua		
	sin agua	9.0%
	agua en el terreno	29.6%
	agua en la vivienda	61.3%

En lo referente a las características de las viviendas se observa que la gran mayoría de las mismas son propias, sin embargo el 13.5 por ciento de las mismas son rentadas, implicando a quienes las habitan cierto gasto. Las frecuencias relativas de los materiales con que está construido el techo muestran que alrededor de un 30 por ciento de las viviendas viven bajo la protección de materiales endebles. En lo relativo a los pisos de las viviendas se observa que alrededor de una décima parte de los hogares tienen pisos de tierra, misma proporción que en el caso de los hogares sin servicio sanitario, lo cual nuevamente puede llegar a asociarse con condiciones de insalubridad.

Características de la vivienda - Posesión de la vivienda		
Posesión de la vivienda		
	Prestada	10.2%
	Recibida como prestación	0.5%
	Rentada o alquilada	13.5%
	Propia y la están pagando	4.9%
	Propia en terreno de asentamiento irregular	0.8%
	Propia en terreno ejidal o comunal	12.8%
	Propia y totalmente pagada en terreno propio	57.2%
	Otro tipo de tenencia	0.2%

Características de la vivienda - Material de los techos		
Material de los techos		
	Cartón, hule, tela, llantas, etc.	0.1%
	Lámina de cartón	3.8%
	Palma, tejamanil o madera	5.0%
	Lámina metálica, fibra de vidrio, plástico o mica	11.4%
	Carrizo, bambú y terrado	1.0%
	Lámina de asbesto	8.2%
	Teja	4.6%
	Panel de concreto	0.5%
	Concreto monolítico	0.9%
	Tabique, ladrillo, tabicón o loza de concreto	59.8%
	Block	0.2%
	Vigueta y poliuretano, viguieta y bovedilla, vigueta y cuña	4.3%

Características de la vivienda - Material del piso		
Material del piso		
	Tierra	9.8%
	Cemento o firme	53.5%
	Madera, mosaico, loseta de concreto, loseta de plástico	36.7%

Características de la vivienda - Servicio Sanitario		
Servicio Sanitario		
	Hoyo negro o pozo ciego	3.2%
	Letrina	10.1%
	Excusado	78.7%
	No dispone de servicio sanitario	8.0%

Es interesante observar que enseres que en la actualidad parecen indispensables aún no se encuentran disponibles para toda la población.

Posesión de enseres	
Enser	
Radio	31.7%
Grabadora	53.8%
Tocadiscos	43.1%
Radiograbadora	89.6%
Televisión	89.7%
Videocasetera	35.0%
Computadora	10.5%
Ventilador	51.3%
Estufa de gas	88.4%
Estufa de otro combustible	4.2%
Refrigerador	74.2%
Licuadora	81.3%
Lavadora	53.3%
Calentador de gas	41.4%
Calentador de otro combustible	2.9%
Vehículo	33.0%
Teléfono	39.6%

En el caso particular del combustible a utilizar, se puede observar que un 14.4 por ciento de los hogares usan leña, esto es un problema que va más allá de la simple incomodidad. Por una parte el uso de leña implica la necesidad de talar árboles, lo cual puede ser perjudicial para las condiciones ambientales del entorno en que viven los hogares. Por otra parte, la población que utiliza la leña como combustible tiene la necesidad de cargarla distancias en ocasiones considerables, afectándoles fisiológicamente, aunado a ello, está la exposición y el riesgo que conlleva el respirar el humo producido por la quema de la leña.

Combustible que utiliza	
Tipo de combustible	
Leña	14.4%
Carbón	0.2%
Petróleo	0.0%
Electricidad	0.4%
Gas	84.0%
Otros	0.1%
No utiliza combustible	1.0%

Cuando se analiza un hogar como unidad desde el punto de vista social y económico, y sin olvidar que en México en este momento aún se conserva el concepto de que el hogar es la unidad mínima de asociación humana, no se puede dejar de lado el analizar las características del individuo que en dicha unidad es considerado como jefe. La importancia de dicho análisis se magnifica en el caso del interés que tiene el presente documento debido a que coincide mayoritariamente que el individuo considerado como jefe del hogar es quien provee de los recursos de subsistencia del mismo. Se tiene que un 18.4 por ciento de los hogares señalan a un individuo de sexo femenino como jefe, así como un 11.8 por ciento de jefes de los hogares que no saben leer ni escribir.

Características del jefe del hogar	
Características	
Jefatura femenina	18.4%
Jefatura analfabeta	11.8%

El acceso a seguridad social implica que los individuos del hogar no tienen que destinar gastos hacia algunas cuestiones imprevistas como enfermedades o accidentes. Menos de la mitad de los hogares del país tienen acceso a este servicio.

Prestaciones sociales	
Acceso a seguridad social	43.5%

Es momento de analizar las variables cuantitativas por medio de sus medias y sus desviaciones estándar. Previamente ya se había mencionado la intención de utilizar indicadores indirectos, esto es, indicadores que combinan la información de dos o más indicadores directamente captados por la encuesta. Ahora se presentan tanto indicadores directos como indicadores indirectos que se considera tienen importancia para alcanzar el objetivo planteado.

Medias de algunos indicadores directos e indirectos		
Variable	Media	Desviación Estándar
Indicadores directos		
Número de cuartos para dormir	2.0	0.9
Total de miembros del hogar	4.3	2.0
Total de hombres	2.1	1.3
Total de mujeres	2.2	1.3
Años de escolaridad del jefe del hogar	8.3	4.4
Ingreso mensual del hogar	6,825.3	10,551.5
Indicadores indirectos		
Índice de hacinamiento (personas por cuarto)	1.9	1.5
Índice de dependencia demográfica (personas en los grupos de edad 0 a 15 años y 65 años en adelante, entre personas en el grupo de 16 a 64 años)	0.7	0.7
Ingreso mensual per cápita del hogar	1,967.4	4,268.2
Ingreso por perceptor	4,696.1	8,833.9
Gasto mensual percapita	1,643.6	3,435.7

En primer lugar se observa el número de cuartos para dormir, que en promedio es de 2 con una desviación de 0.9. Aplicando el Teorema Central del Límite es posible plantear que la distribución de la variable "número de cuartos para dormir" es Normal, con parámetros estimados a partir de la media y la varianza muestrales, con lo que un intervalo de confianza al 95 por ciento para los valores de la variable es el conformado por un centro en la media muestral, y un desplazamiento de dos desviaciones estándar hacia la izquierda y hacia la derecha de la misma. Así, para los cuartos para dormir se tiene un intervalo entre 0.2 y 3.8 cuartos.

El promedio de miembros en el hogar es de 4.3, lo que indica que a nivel nacional la configuración demográfica de los hogares ya se está aproximando al ideal planteado por la política demográfica del gobierno, de conformar familias con una estructura de dos adultos y dos menores, sin embargo la desviación estándar nos señala aún una amplia dispersión, generando un intervalo de 0.1 a 8.5 personas. Es interesante observar que existe una ligera mayoría de mujeres, la cual al ser analizada por medio de una prueba de comparación de medias resulta ser estadísticamente significativa por muy alto margen (mas allá del 99 por ciento de confianza).

Un indicador indirecto que se muestra conveniente construir es el índice de hacinamiento, correspondiente al número promedio de personas que habitan un cuarto. En este caso, dicho indicador corresponde a 1.9 personas por cuarto.

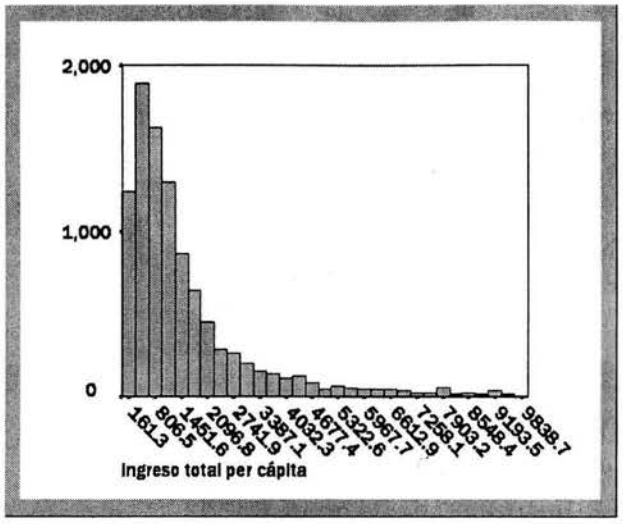
Los años de escolaridad del jefe reflejan generalmente la capacidad de obtener cierto nivel de remuneraciones en el trabajo, pues a mayor escolaridad generalmente se obtendrán mayores ingresos. Esto se constata con el cálculo del coeficiente de correlación de Pearson, que resulta ser de 0.29, positivo y estadísticamente significativo a un nivel de confianza del 99 por ciento. Se tiene un promedio de 8.3 años de escolaridad para los jefes de los hogares, lo cual implica que casi han concluido su educación básica (correspondiente a 9 años), sin embargo la desviación estándar es relativamente grande, generando un intervalo que va desde 0 hasta 17 años, valores casi coincidentes con el valor mínimo y máximo de 0 y 16 años de escolaridad respectivamente.

Otro indicador de interés es el denominado índice de dependencia demográfica, que mide la máxima capacidad que tiene un hogar de distribuir la carga del sustento por medio del cociente entre el número de personas que no están en edad de trabajar entre el número de personas que sí lo están. El valor de dicho índice tiene una media de 0.7, lo cual indica que a nivel nacional cada persona que se encuentra en edad de trabajar tiene que sostener prácticamente a otra persona.

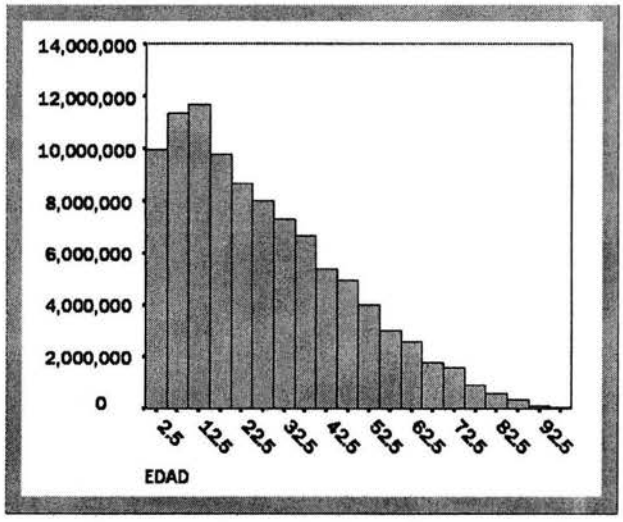
Los indicadores sobre el ingreso y el gasto son indispensables en el análisis. Para el ingreso por perceptor se tiene una media de 4,696.1 pesos mensuales pero con una muy alta desviación estándar de 8,833.9, lo cual remarca la amplia dispersión en la distribución del ingreso.

Es interesante comparar los ingresos per cápita con los gastos per cápita, puesto que reflejan la capacidad de ahorro de los hogares. En este caso la diferencia de las medias es de alrededor de 320 pesos mensuales per cápita (estadísticamente significativa al 99 por ciento de confianza), cantidad que en promedio podrían destinar al ahorro los hogares.

Otro análisis de interés es la construcción de los histogramas para los ingresos y para las edades, lo cual nos mostrará los sesgos existentes en las distribuciones de ambos indicadores. Como se puede apreciar en el histograma siguiente, la distribución del ingreso per cápita tiene un sesgo a la derecha, es decir, la mayor parte de los individuos tiene ingresos bajos. De hecho la mediana es de 1,033.1 pesos, lo cual implica que la mitad de los hogares tienen un ingreso per cápita menor a dicho monto.



En cuanto a la edad se observa que la mayor parte de la población se encuentra en edades jóvenes, y que en los años recientes está reestructurándose la distribución por la dinámica demográfica.



Este es en general el perfil de los hogares en México en el año 2000, como puede apreciarse no todos los individuos tienen acceso a los servicios o bienes que en la actualidad se consideran como necesarios para desarrollar una vida plena.

Capítulo II. Propuestas Históricas de Solución al problema de clasificación de la población por su condición de pobreza

a. El problema de clasificación de la población por su condición de pobreza

Para obtener una agrupación de la población de acuerdo con su condición de pobreza y posteriormente utilizarla para generar la regla de clasificación se requieren establecer ciertos puntos que a continuación se enumeran (Comité Técnico de medición de la Pobreza, 2002). 1) El primero es una definición del concepto de pobreza, esto es, establecer el significado del término. Esta definición tiene un carácter subjetivo, puesto que plantea el hecho de tener un nivel de vida mínimo deseable, y con ello la problemática de establecer dicho nivel mínimo así como los indicadores que lo establecen.

2) En segundo lugar se encuentra la necesidad de establecer la unidad de análisis, es decir, en qué nivel de desagregación se realizará la clasificación. Puede ser a nivel de los individuos, o a nivel de los hogares dadas las características intrínsecas de la sociedad en cuanto a su estructura de asociación colectiva, e incluso puede ser a nivel de agrupaciones geográficas dependiendo también de los objetivos finales de la agrupación, lo que conduce al concepto de marginación en lo colectivo.

3) Un tercer punto estriba en ubicar la fuente de información o en su caso de generarla, esto es en otros términos, identificar una encuesta o censo que contenga la información necesaria para realizar la cuantificación de los niveles de vida, o preferentemente, diseñar una específicamente para tal efecto.

4) Una vez que se dispone de la fuente de información un cuarto punto consiste en la necesidad de identificar las variables para medir el concepto de pobreza de acuerdo con la definición planteada.

5) Finalmente, se procede a aplicar la metodología seleccionada para realizar la agrupación de la población de acuerdo con su condición de pobreza y posteriormente utilizarla para generar la regla de clasificación.

Como se puede apreciar, existe un gran número de combinaciones posibles, derivadas tanto de las opciones a elegir como de la subjetividad que puede implicarse, esto provoca diferencias entre las agrupaciones generadas reflejando la complejidad del problema.

El primero y el último de los puntos establecidos conforman lo que se conoce como problema de medición de la pobreza, tema ampliamente estudiado en los últimos años por investigadores de todo el mundo, de donde han surgido diversas propuestas de solución con diferentes enfoques.

b. Propuestas históricas de la medición de la pobreza

i) Enfoque del Ingreso o del Gasto

Una primera alternativa de carácter unidimensional para medir la pobreza establece como definición de pobreza la posibilidad de realizar un consumo de bienes mínimo para tener un nivel de vida aceptable, utilizando para ello como metodología la comparación de una medida de consumo (gasto) con un gasto mínimo definido como "línea de pobreza". (Ravallion, 1996). Una ventaja de este método de medición es que se basa en un solo indicador por lo que es simple de aplicar. Sin embargo, esto se puede volver también una desventaja cuando la captación se realiza de manera poco confiable. Otra ventaja consiste en que la medición realizada por esta metodología permite realizar comparaciones entre individuos para saber qué individuo es más pobre, así como poder tener medidas de profundidad y severidad de pobreza, entendidas como la lejanía que en conjunto tienen los pobres de la línea de pobreza.

Las desventajas se pueden separar principalmente en dos rubros, el primero asociado con el problema de obtener el valor o monto del gasto que servirá como punto de comparación, para lo cual se tiene que definir claramente el concepto de condición de vida "aceptable", lo cual genera controversia por la subjetividad que conlleva. El segundo rubro se relaciona con los problemas de captación del gasto, como el costo o la falta de declaración que impiden una correcta medición. Un problema adicional consiste en que esta medición es unidimensional puesto que se basa solo en un indicador, esto se menciona debido a que dadas las preferencias de consumo de la gente, al medir la pobreza por medio del gasto se pueden cometer errores que se podrían reducir si se midiera un mayor número de indicadores. Derivada de esta alternativa y con el objetivo de solventar las desventajas y limitaciones señaladas, algunos autores han propuesto la utilización del ingreso en sustitución del gasto, esto debido a una mayor sencillez en su captación.

ii) Enfoque de las Necesidades Básicas Insatisfechas

Una segunda alternativa de tipo multidimensional para la medición de la pobreza es el concepto de las Necesidades Básicas Insatisfechas, planteando una medición de la pobreza a partir de la definición de un conjunto de indicadores convenidos como necesarios para tener una condición de vida "aceptable", y como metodología de clasificación, la determinación de que un hogar es pobre cuando no tiene alguno de los indicadores convenidos. La medición de la pobreza utilizando el enfoque de las Necesidades Básicas Insatisfechas utiliza como indicadores los relativos a enseres y características de las viviendas principalmente. (Comité Técnico de medición de la Pobreza, 2002). Como ventaja de esta alternativa se puede señalar el uso de más de un indicador para la medición de la pobreza. Una desventaja es nuevamente la problemática de captación, que incluye el costo, la confiabilidad, y la falta de declaración de los indicadores utilizados. Otra desventaja radica en que la selección de los indicadores es subjetiva. Finalmente una limitación notoria de este método es que no permite realizar un ordenamiento claro de la población, esto es, no permite decir de manera natural si un individuo es más pobre que otro, a menos que haya algún orden en los indicadores.

iii) Enfoque de los Funcionamientos y las Capacidades o Potencialidades

Un tercer enfoque también multidimensional consiste en analizar los conceptos de Funcionamientos y Capacidades o Potencialidades. (Sen, 1997). Los Funcionamientos se refieren a cuestiones como acceso a alimentación adecuada y suficiente, apropiado vestido y cobijo, así como integración social y acceso a los beneficios de desarrollo, es decir, se refieren a cuestiones que definen una condición en el momento en que son medidos. Las Capacidades y Potencialidades por su parte se refieren a aspectos que tienen que ver con la oportunidad que tendrán los individuos de alcanzar los funcionamientos por ellos mismos en un momento posterior, como pueden ser la educación o la capacitación laboral. El uso del enfoque de Funcionamientos y Capacidades o Potencialidades para realizar la medición de la pobreza utiliza además de los indicadores usados por el enfoque de las Necesidades Básicas Insatisfechas, que corresponden en parte a los Funcionamientos, otros indicadores como la escolaridad de los individuos o el acceso a servicios de salud entre otros.

La ventaja principal del uso de estos conceptos para medir la pobreza radica en que toma en cuenta el posible bienestar futuro de los individuos. Las desventajas son los ya señalados problemas en la captación además de la definición en los aspectos relativos tanto a los Funcionamientos como a las Capacidades o Potencialidades.

iv) Otros enfoques multidimensionales

Se han propuesto otros enfoques multidimensionales partiendo de definiciones variadas como oportunidades de participación política, o de mecanismos de apropiación de recursos, o de bienestrar subjetivo, todos ellos con ventajas y desventajas. Propuestas metodológicas como la medición de índices de pobreza humana, mediciones mediante lógica difusa (análisis que contempla que un individuo pueda no estar asignado a un grupo debido a la imposibilidad de determinar claramente su agrupación a partir de sus características) o la utilización de conceptos de distancia se plantean como alternativas menos exploradas, así como aquéllas metodologías utilizando técnicas estadísticas. (López-Calva y Rodríguez Chamussy, 2004).

Es importante señalar que todas estas propuestas de medición de la pobreza multidimensional conllevan a problemas que cuando no pueden ser solventados, no conducen a solucionar el problema de agrupación de la población por su condición de pobreza. Estos problemas son la determinación del tipo de información que se requiere para poder llegar a medidas multidimensionales, el tipo de dimensiones que son relevantes, y la interacción entre ellas evitando multicolinearidad.

Existen enfoques mixtos, que utilizan varias de las propuestas referidas previamente. Sería conveniente buscar una estrategia para combinar todos los enfoques, de manera que la medición de la pobreza tuviera un enfoque integral.

c. Aplicación de algunas propuestas históricas de la medición de la pobreza en el caso de México

i) Enfoque del Ingreso o del Gasto

Como se ha mencionado, el enfoque del ingreso o del gasto se sustenta en que existe un monto de ingreso o de consumo que refleja un nivel mínimo de "bienestar", es decir, que es suficiente para cubrir sus necesidades mínimas que pueden englobar el concepto de alimentación, de otros gastos como educación o salud, además de gastos en vestido, calzado, vivienda y transporte.

Es importante señalar que la definición del concepto de pobreza dependerá de la situación particular que se quiere analizar. En el año 2002 se reunió a un grupo de expertos en la materia, con el objetivo de obtener por medio del consenso una medida de pobreza que pudiera ser llevada a la práctica.

Después de realizados los trabajos de análisis, los resultados de dicho consenso consistieron en determinar que el enfoque de ingreso o de gasto era la mejor opción para realizar la medición de la pobreza. Para fines de la clasificación, se definió al hogar como unidad de análisis, y se obtuvieron tres niveles de ingreso o gasto mínimo necesario dependiendo de tres distintas definiciones de pobreza que a continuación se describen:

1) Línea de pobreza alimentaria: definida como el ingreso mínimo necesario para cubrir las necesidades de alimentación, que en monto corresponde a 15.4 pesos diarios[6] por persona en zonas rurales (de menos de 2,500 habitantes) y a 20.9 pesos diarios por persona en las zonas urbanas. De acuerdo a la ENIGH del año 2000 se tiene que 18.6 por ciento de los hogares (24.2 por ciento de la población) tiene un ingreso inferior a esta línea de pobreza alimentaria.

2) Línea de pobreza de capacidades: definida como el ingreso mínimo necesario para cubrir, además de las necesidades de alimentación, los gastos en educación y salud, que en monto corresponde a 18.9 pesos diarios por persona en zonas rurales y a 24.7 pesos diarios por persona en zonas urbanas. De acuerdo a la ENIGH del año 2000 se tiene que 25.3 por ciento de los hogares (31.9 por ciento de la población) tiene un ingreso inferior a esta línea de pobreza de capacidades.

3) Línea de pobreza de patrimonio: definida como el ingreso mínimo necesario para cubrir las necesidades de alimentación, los gastos en educación y salud, así como los gastos en vestido, calzado, vivienda y transporte, que en monto corresponde a 28.1 pesos diarios por persona en zonas rurales y a 41.8 pesos por persona en zonas urbanas. De acuerdo a la ENIGH del año 2000 se tiene que 45.9 por ciento de los hogares (53.7 por ciento de la población) tiene un ingreso inferior a esta línea de pobreza de patrimonio.

Es importante señalar que las definiciones conllevan un carácter de subjetividad, puesto que determinar un nivel mínimo de necesidades no es simple. Esto se trata de resolver en el caso alimentario por medio de una canasta básica de bienes que incluyan los nutrientes recomendados para una sana alimentación. Para los casos que involucran cuestiones no alimentarias es más complicada la definición, puesto que la subjetividad juega un rol de mayor trascendencia. Surgen preguntas naturales, como determinar cuánto ingreso es necesario para cubrir las necesidades de educación o de salud. En este punto se puede apreciar nuevamente y con claridad la complejidad del problema.

[6] Pesos del año 2000

ii) Enfoque de las Necesidades Básicas Insatisfechas

El enfoque de Necesidades Básicas Insatisfechas depende directamente de la definición de las Necesidades que son Básicas. Por ejemplo, si se definiera que la disponibilidad de agua, de luz y de drenaje son necesidades básicas, entonces un 26.8 por ciento de los hogares en el año 2000 tenían insatisfechas dichas necesidades básicas.

Es necesario insistir en que en este enfoque no es clara la determinación de un orden, esto es que no se puede identificar claramente entre dos hogares pobres quién lo es más, o qué tan lejos están entre sí o de la no pobreza.

iii) Otros enfoques

El enfoque de las capacidades y potencialidades ha sido utilizado por el programa de combate a la pobreza más importante del gobierno federal a finales del siglo XX y principios del XXI, el Programa de Desarrollo Humano Oportunidades, el cual se basa en una definición de la pobreza que contempla tanto la posesión de capital físico (características de las viviendas y posesión de enseres) como de capital humano (composición demográfica del hogar, escolaridad, situación laboral, etc.).

En un trabajo realizado en el año 2004 por investigadores de la Universidad de las Américas, (López-Calva y Rodríguez Chamussy, 2004), se discute la posibilidad de utilizar otros enfoques multidimensionales como el índice de pobreza humana o el enfoque de lógica difusa (conjuntos difusos). El trabajo busca comparar los resultados de estas técnicas con los planteados por el Comité Técnico de Medición de la Pobreza, encontrando como conclusión que no existen ventajas sustantivas en la utilización de los métodos multidimensionales mencionados, por la razón de que no existen dentro de la fuente de información datos relativos a desnutrición, riesgos de salud, ansiedad e inseguridad, falta de participación política, discriminación y baja autoestima, los cuales son elementos que para los autores definen varias dimensiones de la pobreza desde la visión multidimensional.

Capítulo III. Técnicas estadísticas de solución al problema de agrupación

a. Indicadores y métodos de selección

Para aplicar las técnicas estadísticas de agrupación es necesario seleccionar de entre todos los indicadores disponibles, algunos que servirán para obtener resultados confiables. Esto debido a que hay indicadores redundantes, es decir, que proporcionan la misma información. Por ello, se debe tratar de combinar la información disponible para simplificarla al máximo, intentando por una parte reducir en la medida de lo posible las categorías de las variables cualitativas, y por otra tratando de identificar aquéllas variables que proporcionen mayor información para realizar la agrupación.

Existen diversas técnicas estadísticas que permiten identificar los indicadores de mayor utilidad para solucionar el problema de agrupación. Entre ellas se pueden mencionar, para variables de tipo discreto a las medidas de asociación y al análisis de correspondencias, y para variables continuas a los métodos de regresión. A continuación se describen brevemente dichas técnicas.

i) Medidas de Asociación

Las variables categóricas pueden o no tener una relación entre sí, y una forma de cuantificar dicha relación consiste en aplicar una prueba de hipótesis para definir si existe independencia entre dos variables. Dicha prueba de hipótesis se basa en la estadística χ^2 la cual es el resultado de la suma de las diferencias elevadas al cuadrado entre las frecuencias observadas y las esperadas bajo el supuesto de independencia, estandarizadas por las frecuencias esperadas bajo el supuesto de independencia. La fórmula para su cálculo se presenta a continuación.

$$\chi^2 = \sum \frac{(obs - esp)^2}{esp} .$$

La estadística χ^2 tiene una distribución asintótica de tipo χ^2 con k grados de libertad, donde k es el número de frecuencias que se requieren conocer en la tabla de frecuencias para determinarla completamente dependiendo del esquema de muestreo. La hipótesis nula bajo el supuesto de independencia consiste en plantear que las frecuencias observadas y las esperadas son iguales y por tanto la estadística χ^2 será igual a cero.

Una vez que se ha comprobado la existencia de asociación entre las variables de tipo categórico, el siguiente paso consiste en cuantificar el nivel de relación que tienen entre sí las variables, ya que esto permitirá definir cuáles de dichas variables serán de mayor utilidad para realizar la agrupación. Algunas medidas de asociación parten de la estadística χ^2, entre dichas medidas se encuentra la denominada fi-cuadrada (ϕ^2), que es simplemente el cociente de la estadística χ^2 entre el total de individuos en estudio y que se calcula para normalizar la estadística χ^2. A partir de la medida (ϕ^2) se obtiene el Coeficiente de Contingencia en Media Cuadrática de Pearson que se calcula como:

$$P = \sqrt{\frac{\phi^2}{\phi^2 + 1}} \, .$$

El Coeficiente de Contingencia en Media Cuadrática de Pearson toma valores entre cero y uno. Cuando el coeficiente de Contingencia en Media Cuadrática de Pearson vale cero es equivalente al hecho de que las variables sean independientes, y mientras mayor sea su valor, mayor nivel de relación de dependencia existirá entre las variables.

ii) Análisis de Correspondencias Simples

El Análisis de Correspondencias Simples es una técnica estadística que permite reflejar de manera gráfica la asociación que tienen las categorías de dos variables de tipo cualitativo (Lebart, Morineau & Warwick, 1977). La técnica consiste en plantear la tabla de contingencia (frecuencias) como una matriz con vectores renglón y vectores columna, y generar las "mejores" proyecciones ortogonales de los vectores renglón y vectores columna con respecto a vectores unitarios que tienen relación directa entre sí, de modo que es posible llevar mediante ecuaciones de transición, a las proyecciones de los vectores renglón al espacio de las proyecciones de los vectores columna y así poder graficarlas simultáneamente. A continuación se detalla el procedimiento.

Aunque es posible analizar las frecuencias absolutas de las variables para obtener el análisis, es recomendable utilizar las frecuencias relativas condicionales (distribución de una variable condicional a cada categoría de la otra variable) para evitar posibles sesgos al tener categorías con frecuencias muy grandes o muy pequeñas. Las frecuencias relativas condicionales pueden calcularse por renglón o por columna, con lo cual se tienen dos posibles matrices. Los renglones de la matriz de frecuencias relativas condicionales por renglón se llaman perfiles por renglón, y los renglones de la matriz de frecuencias relativas condicionales por columna se llaman perfiles por columna.

La forma matricial de calcular estas matrices es la siguiente: si se denota por F a la matriz de frecuencias, y por D_r y D_c a las matrices diagonales que contienen como elementos de la diagonal a las frecuencias marginales (frecuencias totales de cada categoría de una variable independientemente de la otra variable) por renglón y por columna respectivamente, entonces la matriz resultante de la operación $D_r^{-1}F$ contendrá los perfiles por renglón, mientras que $D_c^{-1}F'$ contendrá los perfiles por columna.

El objetivo del Análisis de Correspondencias Simples es entonces, encontrar las proyecciones ortogonales óptimas para los perfiles por renglón y por columna y representarlos en el mismo espacio vectorial. Esto se consigue encontrando ejes distintos a los originales, para los cuales la distancia en conjunto de los perfiles hacia dichos ejes sea mínima. En este punto se tiene que destacar que se utiliza la distancia χ^2 en lugar de la euclidiana ya que esta segunda distancia tiene la propiedad de "equivalencia distribucional", que implica que al juntar dos categorías con idéntico perfil, las distancias entre las restantes categorías permanecen invariantes, lo cual no ocurre con la distancia euclidiana. La distancia χ^2 consiste en la estandarización de la distancia euclidiana por el recíproco de las frecuencias marginales (por renglón o por columna según sea el caso).

Para encontrar las proyecciones para los perfiles por renglón se tiene que proyectar a los r vectores conformados por los renglones de la matriz $D_r^{-1}F$ en un vector u que cumple con que $u'D_c^{-1}u = 1$, es decir que es unitario en el sentido de la distancia χ^2. Entonces el vector de las r proyecciones estará dado por $v = D_r^{-1}FD_c^{-1}u$. El proceso de optimización consistente en minimizar las distancias de los puntos originales a sus proyecciones es equivalente a maximizar la suma de cuadrados ponderada de las proyecciones $v^T D_r v$ sujeto a que $u'D_c^{-1}u = 1$, de donde por métodos de álgebra lineal se encuentra que un conjunto de soluciones para u lo conforman los vectores propios de la matriz $F'D_r^{-1}FD_c^{-1}$. El ordenamiento generado por los valores propios correspondientes de forma descendente dictamina las mejores proyecciones.

Si bien el análisis se puede replicar para encontrar las proyecciones de los perfiles por columna, existe una regla directa que permite obtenerlos a partir de los resultados previos. Si denotamos a $\alpha = D_c^{-1}u_i$, como el vector que genera las proyecciones para los perfiles por renglón correspondiente al valor propio λ_i, entonces $\beta = \frac{1}{\sqrt{\lambda_i}}D_r^{-1}F\alpha$ es el vector que genera las proyecciones para los perfiles por columna. Esta regla garantiza la posibilidad de obtener una representación simultánea de los perfiles en un mismo gráfico, mismo que podrá construirse en una, dos y hasta tres dimensiones dependiendo del número de valores propios que se decida usar (Para una discusión más amplia ver Soto, 1999).

iii) Regresión Lineal Múltiple

La regresión lineal múltiple es una técnica estadística que tiene como objetivo establecer la relación lineal existente entre un conjunto de variables denominadas predictores o independientes y una variable de respuesta o dependiente.

El modelo de regresión lineal múltiple se establece de la siguiente manera:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_n X_{pi} + \epsilon_i \ ,$$

donde:

Y_i es la variable dependiente para el i-ésimo individuo;

X_{ji} es la j-ésima variable independiente para el i-ésimo individuo;

β_j es la constante asociada a la variable j-ésima que sirve para establecer la asociación lineal buscada;

Los supuestos que se establecen para el modelo de regresión lineal múltiple son:

α) Los ϵ_i's son mutuamente independientes e idénticamente distribuidos con $\mathrm{E}(\epsilon_i)=0$, $\mathrm{V}(\epsilon_i)=\sigma_\epsilon^2$.

β) La distribución de ϵ es independiente de la distribución conjunta de $X_1, X_2, \ldots, X_n$.

γ) $\beta_0, \beta_1, \ldots, \beta_p$ son constantes desconocidas.

La estimación de las β's se puede realizar por el método de mínimos cuadrados, y hasta este punto no se necesitan supuestos adicionales. Sin embargo para realizar la validación de la significancia estadística de las β's se requiere el supuesto adicional de Normalidad de los ϵ_i's (y por lo tanto de los Y_i's). Con este supuesto se puede establecer la significancia estadística de las β's por medio de pruebas t de Student.

Para validar el modelo en su conjunto se pueden utilizar pruebas F que involucran a la estadística R^2, así como la estadística de Press consistente en correr el modelo dejando fuera una observación a la vez y verificando que los resultados sean invariantes. Para mejorar el modelo se pueden utilizar técnicas de selección de variables como lo son los

métodos de inclusión, los de exclusión o los de inclusión y exclusión. Otra estadística de utilidad para la selección de variables es la Cp de Mallow que mide la mejora en el modelo de acuerdo a las distintas combinaciones posibles de variables a incluir (Jobson, 1992).

Existen varios problemas que pueden conducir a un modelo de regresión inválido, como lo son la multicolinealidad, que se refiere a tener variables altamente correlacionadas. Para identificar esta multicolinealidad se pueden analizar las correlaciones o una estadística conocida como el factor de inflación de la varianza. (Jobson, 1992). Para corregir el problema de la multicolinealidad se pueden omitir variables redundantes, aumentar datos a la muestra (cuando es posible) o mejor aún juntar variables redundantes por medio de alguna técnica estadística de reducción de variables.

Otro problema prevaleciente en las regresiones es el de los valores atípicos (outliers) y los valores influyentes, mismos que pueden sesgar los resultados, por tanto hay que identificarlos y verificar si están mal captados o si existe alguna razón para excluirlos del análisis.

Cuando los supuestos no se cumplen, lo cual se puede verificar por medio de gráficos de residuales, se tiene que buscar alternativas de solución como pueden ser las transformaciones. Lo mismo ocurre cuando el supuesto de linealidad no se cumple, con lo cual el modelo de regresión que se obtenga no sería válido.

iv) Análisis de correspondencias múltiples

Un análisis que es útil para observar de manera gráfica la asociación entre más de dos variables de tipo categórico es el denominado análisis de correspondencias múltiples (Lebart, Morineau & Warwick, 1977), el cual es la extensión multidimensional del análisis de correspondencias simples.

El análisis de correspondencias múltiples parte de las mismas bases del análisis de correspondencias simples en términos de buscar la "mejor" representación gráfica de las categorías de las variables por medio de los conceptos de proyecciones ortogonales. Se generaliza a partir de la demostración de la equivalencia entre realizar un análisis de correspondencias simples a la tabla de frecuencias y realizarlo a la tabla donde se enlista a los individuos de manera separada, utilizando variables indicadoras para señalar a qué categoría pertenece cada individuo. Es claro que bajo el segundo esquema no existe una limitante del número de variables indicadoras a incluir en el análisis, por lo que más de dos variables pueden ser consideradas.

Existe además la posibilidad de probar que es equivalente analizar la matriz de variables indicadoras que la forma cuadrática de la misma en lo que se conoce como tabla de Burt, misma que tiene una dimensión considerablemente menor que la matriz de variables indicadoras simplificando los cálculos numéricos (Para conocer la demostración ver Soto, 1999).

b. La agrupación inicial y el algoritmo de "Simulated Annealing"

El planteamiento original que se hizo del problema de clasificación involucra la intención de identificar el grupo al cual pertenece un individuo de una población de acuerdo con un conjunto de características, estas condiciones se especificarán más detalladamente cuando se describan posteriormente las propuestas de solución, sin embargo se menciona a la agrupación inicial en este punto, porque en la propuesta bayesiana de solución al igual que en el análisis discriminante o en la regresión logística se parte del supuesto de que para un conjunto de individuos de la población se conoce el grupo al que pertenece, y a partir de dicha agrupación y de las características de los individuos se construye una regla de clasificación para nuevos individuos para los que no se conoce el grupo al que pertenece.

De hecho, el conjunto de individuos para el que se conoce su agrupación sirve como base para evaluar si la regla de clasificación funciona de manera conveniente. Sin embargo, en muchas situaciones (como en el problema que interesa a este trabajo) no se tiene una definición exacta de la agrupación de los individuos que servirán para generar la regla, por lo que es necesario explorar métodos para encontrar primero una agrupación adecuada.

En particular el problema que interesa al presente trabajo no tiene una agrupación inicial clara. Se tomarán dos alternativas para obtener la agrupación inicial deseada. Es importante recordar que dentro de los objetivos del presente proyecto se encuentran por una parte evaluar la propuesta oficial de agrupación, y por otra tratar de encontrar una mejor propuesta si es posible.

Dados los objetivos planteados, la primera alternativa de agrupación inicial, será justamente la obtenida por el Comité Técnico para la Medición de la Pobreza en México, consistente en obtener puntos de corte por ingresos per cápita, generando cuatro grupos bien definidos. El de aquéllos hogares que no logran cubrir los gastos en alimentación; el de los hogares que si bien cubren los gastos de alimentación no lo hacen en lo que respecta a educación y salud; el de los hogares que aunque pueden cubrir los gastos en alimentación,

educación y salud, no logran cubrir el costo de vestido, calzado, vivienda y transporte, y finalmente el cuarto grupo de hogares para quienes su ingreso es suficiente para cubrir todos los rubros previamente mencionados.

Es importante mencionar que el uso de esta alternativa de agrupación inicial correspondiente a la obtenida por el Comité Técnico para la Medición de la Pobreza en México permitirá evaluar la propuesta de regla de clasificación planteada por el Comité (que consiste en aplicar los mismos criterios que los utilizados para generar la agrupación inicial, para un nuevo hogar comparando su ingreso per cápita con una "línea de pobreza"), además de determinar si los indicadores elegidos son adecuados y suficientes para generar una nueva propuesta de regla de clasificación desde una perspectiva multidimensional.

Para generar la segunda alternativa de clasificación inicial se utilizará una técnica denominada como "Simulated Annealing". Dicha técnica es un algoritmo de optimización basado en una analogía con un proceso químico para obtener cristales puros que se denomina "annealing", y que consiste en enfriar paso a paso el material, dando tiempo en cada paso a que la estructura atómica del cristal alcance su mínimo nivel de energía a la temperatura de ese paso (Kirkpatrick, Gelatt & Vechhi, 1983).

La analogía con un proceso de optimización se da en que para una función $f(x)$ que se desea minimizar se parte de un punto aleatorio x_0 y se selecciona en el paso i-ésimo a un punto x_i a una distancia d del punto x_{i-1} del paso previo, eligiendo desplazarse hacia el punto x_i con probabilidad uno si $f(x_i) < f(x_{i-1})$, y en caso contrario eligiendo el punto x_i con probabilidad $exp^{-\delta/t}$, donde $\delta = f(x_i) - f(x_{i-1})$ y donde t es un parámetro (inicialmente grande) que simula la acción de la reducción de la temperatura paso a paso en el proceso químico de cristalización previamente descrito.

El proceso se puede aplicar para maximizar una función invirtiendo la desigualdad en el paso de decisión y comenzando por un valor t pequeño que vaya aumentando paso a paso. La idea de utilizar este procedimiento para optimizar surge de que en ocasiones las funciones que se analizan tienen muchos mínimos y máximos locales. Este procedimiento permite salir de las zonas de valles o cimas con mínimos y máximos locales (Aarts & Korst, 1989).

La aplicación de la técnica para generar una agrupación a partir de un conjunto de datos se realiza de la siguiente manera: la función que se buscará maximizar será la utilidad esperada que tiene una agrupación del conjunto de datos disponible lo cual conducirá a que se obtenga una agrupación "optima", para ello se partirá de una agrupación con un

número de grupos grande y se reducirá de manera iterativa juntando grupos en cada paso de acuerdo con algun criterio.

Así pues, se denotará a $G = \{G_1, G_2, \ldots, G_s\}$ como una partición en s grupos del conjunto de individuos, y a $u^*(G)$ como una función de utilidad esperada de tener dicha partición. Entonces la aplicación del algoritmo de *Simulated Annealing* para reducir el tamaño s de los grupos consiste en:

1) Definir un número inicial grande de grupos.

2) Definir una "temperatura inicial" t_0. (el nombre del término de temperatura se hereda del origen del método).

3) Iterar hasta que el número de veces consecutivas en que la clasificación permanezca igual alcance un máximo previamente definido.

4) Iterar elevando la "temperatura" en cada paso.

5) Definir el número de grupos que se juntan en el siguiente paso y el criterio para juntarlos. El procedimiento de selección de los grupos que se juntan deberá contener un componente aleatorio.

6) Generar la utilidad esperada de la clasificación G_i y de la G_{i+1} (paso i-ésimo) y aplicar la siguiente regla:

Elegir G_{i+1} si $u^*(G_{i+1}) > u^*(G_i)$.

Elegir G_{i+1} si $u^*(G_{i+1}) <= u^*(G_i)$ pero $exp\{\frac{-(u^*(G_i)-u^*(G_{i+1}))}{t}\} < aleatorio(0,1)$.

Elegir G_i en otro caso.

En la aplicación de este trabajo se propone agrupar a los hogares de acuerdo con su nivel de ingresos mensuales per cápita por medio de los centiles, de modo que se empiece con 100 grupos. Debido a la nula información previa sobre el comportamiento de las utilidades, se parte de una temperatura inicial $t_0 = 1$, misma que se incrementará en 20 por ciento cada paso, utilizando además como criterio de paro 10 iteraciones continuas con clasificación invariante. Se plantea reducir de uno en uno el número de grupos, y el criterio para definir la nueva agrupación consiste en escoger un grupo aleatoriamente, y despues escoger a su

vecino mas cercano (existe el concepto de ordenamiento dado que la clasificación inicial se hizo por medio de centiles de ingresos), y unir esos 2 grupos para genera un nuevo grupo, dejando el resto como estaban. La utilidad se plantea como p*log(p) con p siendo la probabilidad de que un individuo sea asignado a un grupo determinado. Los resultados de la aplicación se presentan en el cuarto capítulo.

c. Análisis Discriminante

La propuesta de Fisher para realizar un análisis discriminante cuando se considera una población con más de dos grupos consiste en generar las combinaciones lineales de los vectores de variables de los individuos que mejor discriminan de acuerdo a algún criterio de optimalidad, entre los grupos definidos a priori (Johnson & Wiechern, 1988). Las combinaciones lineales $l'x$ también llamadas funciones lineales discriminantes son:

$$l_j' x_i = a + P_{j1} X_{1i} + P_{j2} X_{2i} + \ldots + P_{jn} X_{ni} \ ,$$

donde:

$l_i' x_j$ es el score discriminante de la j-ésima función para el i-ésimo individuo;

a es una constante;

P_{jk} es el peso discriminante de la función j-ésima para la variable independiente k-ésima;

X_{ki} es la variable independente k-ésima para el i-ésimo individuo;

La derivación de las funciones lineales discriminantes parte de la idea de generar una separación de los grupos, asumiendo igualdad de varianzas, esto es que las matrices de varianzas y covarianzas de las variables en cada grupo son iguales entre sí: $\sum_1 = \sum_2 = \ldots = \sum_g$. Si se denota como $\overline{\mu}$ al promedio de las medias de los grupos y como B_0 a la suma de las distancias al cuadrado existentes entre las medias de los grupos y el promedio de las medias de la siguiente forma:

$$\overline{\mu} = \frac{1}{g} \sum_{i=1}^{g} \mu_i \qquad \text{y} \qquad B_0 = \sum_{i=1}^{g} (\mu_i - \overline{\mu})(\mu_i - \overline{\mu})' \ .$$

Y considerando la combinación lineal $Y = l'X$ se tiene que:

$$E(Y) = l'E(X|p_i) = l'\mu_i$$

para la población p_i, y

$$V(Y) = l'Cov(X)l = l'\sum l$$

para todas las poblaciones.

Por tanto el valor esperado de Y cambia dependiendo de la población de la cual proviene X. La media total para Y se obtendría entonces como:

$$\mu_y = \frac{1}{g}\sum_{i=1}^{g} l'\mu_i = l'(\frac{1}{g}\sum_{i=1}^{g}\mu_i) = l'\overline{\mu}\ .$$

Si se utiliza la razón de la suma de distancias al cuadrado de entre las medias de los grupos y la media total para Y con respecto a la variabilidad "total" medida como la varianza de Y, se obtiene la forma algebraica:

$$\frac{l'B_0 l}{l'\sum l}\ ,$$

entonces, si se maximiza con respecto a l, reflejará la combinación lineal que mejor separa los grupos.

Por medio de descomposición espectral se puede probar que los eigenvectores de la matriz $\sum^{-1} B_0$ corresponden a los coeficientes l que maximizan la razón anteriormente descrita.

Ahora bien, dado que en general no se dispone de valores poblacionales para μ y $\sum$, es necesario utilizar estimadores, de la siguiente forma:

$$\hat{B}_0 = \sum_{i=1}^{g}(\overline{x}_i - \overline{x})(\overline{x}_i - \overline{x})' \qquad \text{y} \qquad \hat{\sum} = W/n_1 + n_1 + \ldots + n_g - g\ ,$$

donde

$$W = \sum_{i=1}^{g}(n_i - 1)S_i = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)(x_{ij} - \overline{x}_i)'\ .$$

En este caso las funciones lineales discriminantes muestrales $\hat{l}'x$ se obtienen encontrando los eigenvectores de la matriz $W^{-1}\hat{B}_0$.

Una vez que se han calculado las funciones lineales discriminantes, el procedimiento de clasificación consiste en asignar x a la población p_z si:

$$\sum_{j=1}^{r}[\hat{l}'(x-\overline{x}_z)]^2 \leq \sum_{j=1}^{r}[\hat{l}'(x-\overline{x}_i)]^2$$

para todo grupo i distinto de z, donde r es el número de funciones discriminantes a utilizar.

El supuesto de Igualdad de Varianzas y Covarianzas entre los grupos se puede validar por medio de pruebas de hipótesis, la forma de corregir este supuesto es utilizando distancias no euclidianas como la de Mahalanobis, que estandariza los datos para obtener el supuesto, otra manera de corregirlo es aumentando el tamaño de muestra cuando es posible y una forma más es utilizando la clasificación por medio de funciones discriminantes no lineales (como la cuadrática por ejemplo). (Huberty, 1994).

Es importante corroborar que la información proporcionada por distintas variables no sea redundante, es decir, que no haya multicolinearidad, y para ello se analiza a la matriz de varianzas y covarianzas. (Hair et. al., 1984). La forma de remediar la presencia de multicolinearidad es la aplicación de métodos de selección de variables, como el de selección hacia adelante incluyendo una variable a la vez, el de selección hacia atrás excluyendo una variable a la vez, o el de selección por pasos sucesivos permitiendo inclusión y exclusión de variables en cada paso. Se requiere además verificar la presencia de outliers.

Los cálculos numéricos para obtener los pesos discriminantes de las funciones se pueden efectuar por medio de algoritmos conocidos de optimización. Una vez obtenidos dichos cálculos, el siguiente paso consiste en evaluar los resultados. La mejor forma de evaluar dichos resultados es comparar la clasificación generada por el análisis con la agrupación inicial, sin embargo existen otras pruebas estadísticas para evaluar el poder discriminatorio, como la lambda de Wilks o la traza de Hotelling. Ambas son pruebas de comparación de medias que permite definir si los valores medios de los grupos que se han definido por la clasificación usando la técnica de discriminante descrita son distintos entre sí estadísticamente.

d. Regresión Logística Multinomial

La regresión logística multinomial se describirá a continuación en el contexto de su aplicación como generadora de una regla de clasificación. La regresión logística multinomial se define como un modelo matemático que describe la relación entre un conjunto de variables cuantitativas $X's$ y una variable categórica. En el contexto de su aplicación como generadora de una regla de clasificación, la regresión logística multinomial permite definir las probabilidades de que un individuo provenga de cada uno de los grupos generados por las categorías de la variable categórica mencionada previamente, y con ello asignar al individuo al grupo para el que la probabilidad sea mayor. A continuación se detalla la forma en que se obtienen dichas probabilidades.

Primero se presenta la función logística de la que parte el modelo de regresión logística multinomial, cuya forma matemática es:

$$f(z) = \frac{1}{1 + exp\{-z\}} \ ,$$

dicha forma funcional tiene entre sus atributos, la característica de que su rango se encuentra en el intervalo (0,1), siendo 0 y 1 sus límites, y correspondiendo a una función continua creciente. Estas características permiten que dicha función se tome como base para modelar los valores de una probabilidad.

Para obtener el modelo de regresión logística multinomial a partir de la función logística se define a z como una combinación lineal de las variables independientes $X's$ y sus correspondientes parámetros β de la siguiente forma:

$$z = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n \ ,$$

donde X_i es la i-ésima variable independiente.

Sustituyendo en la función logística se tiene: $f(z) = \frac{1}{1+exp\{-(\beta_0+\beta_1 X_1+\ldots+\beta_n X_n)\}}$.

Esta función se utilizará para modelar la probabilidad $f(z_k)$ de que un individuo esté en el grupo k-ésimo, para lo cual se tendrán que definir tantas funciones como grupos existan de la siguiente forma:

$$f(z_k) = \frac{1}{1 + exp\{-(\beta_{0k} + \beta_{1k} X_1 + \ldots + \beta_{nk} X_n)\}} \ .$$

Utilizando la transformación conocida como "logit" se obtiene la siguiente expresión:

$$ln(\frac{f(z_k)}{1-f(z_k)}) = \beta_{0k} + \beta_{1k}X_1 + \ldots + \beta_{nk}X_n \ .$$

Es en este punto donde el modelo toma la forma de una regresión, donde los valores de $f(z_k)$ provienen de las proporciones de individuos en cada grupo. Para obtener estimadores para las $\beta's$ se plantea el supuesto de distribución logística y se aplican técnicas como la de máxima verosimilitud. Existen algunas herramientas estadísticas específicas para validar la significancia del modelo de regresión logística que pueden ser aplicados, como la estadística de Wald que es el cociente del valor estimado del parámetro entre su desviación estándar estimada, y que sirve para validar la hipótesis de que el parámetro es estadísticamente distinto de cero. (Kleinbaum, 1992).

e. Una propuesta con enfoque Bayesiano

A continuación se presenta la teoría que sustenta una propuesta con enfoque Bayesiano desarrollada por J. M. Bernardo en 1988 para resolver el problema de Clasificación.

Se busca clasificar a un conjunto de N individuos de una población P en k grupos denotados por $\{G_i : i \in J_k\}$ que forman una partición de la población. Se parte en primera instancia del supuesto de que para un subconjunto de los N individuos, digamos de tamaño n con $n < N$, se tiene un banco de datos constituido por una matriz donde los vectores denotan r características de los individuos, así como el grupo al que pertenecen, esto es: $D = \{\underline{x}_j, G_{(j)} : j \in J_n\}$ donde $\underline{x}_j \in R^r$ es el vector de r atributos o características y $G_{(j)} \in \{G_i : i \in J_k\}$ es el grupo al que pertenece el individuo j.

Para un individuo de la población P con características $\underline{x}$, que no se encuentra en el subconjunto de n individuos que conforman el banco de datos, se busca determinar el grupo al que pertenece. Para ello se propone obtener primero una distribución de probabilidad predictiva que denotaremos $\{p(G_i|\underline{x}, D) :\in J_k\}$, que no es más que la probabilidad de que el individuo pertenezca a cada uno de los k grupos dado su vector de atributos y la información contenida en el banco de datos, para posteriormente asignar el individuo al grupo para el cual la probabilidad referida sea mayor.

Solución desde el punto de vista Bayesiano

Por el Teorema de Bayes se tiene que:

$$p(G_i|\underline{x}, D) \propto p(G_i|D)p(\underline{x}|G_i, D) \text{ para cada } i \in J_k \tag{1}$$

con la constante de proporcionalidad dada por:

$$\frac{1}{p(\underline{x}|D)} .$$

Donde $p(\underline{x}|G_i, D)$ son los modelos que describen el comportamiento de los atributos dado el banco de datos D para cada grupo, y $p(G_i|D)$ es la distribución inicial de cada grupo.

Para obtener $p(G_i|\underline{x}, D)$ es claro a partir de la expresión (1), que se requiere especificar

$$\{p(G_i|D), p(\underline{x}|G_i, D) : \underline{x} \in R^r\} \text{ para cada } i \in J_k .$$

La especificación de $p(G_i|D)$ se realiza a partir del supuesto de que dicha probabilidad dependerá de D sólo por el diseño muestral y no de los valores específicos de $\underline{x}_j : j \in J_n$. En caso de un muestreo retrospectivo $p(G_i|D)$ corresponderá a la distribución generada por el tamaño de los grupos, mientras que en caso de un muestreo prospectivo corresponderá a una distribución uniforme.

Por otra parte, para la especificación de $p(\underline{x}|G_i, D)$ se plantean tres supuestos:

A) Existe una transformación $t : R^r \to R^l$ (para alguna $l \in N$) tal que

$$p(G_i|\underline{x}, D) \approx p(G_i|t(\underline{x}), D) \text{ para cada } i \in J_k ,$$

es decir que tiene t es aproximadamente una estadística suficiente. La prueba de que este supuesto se puede cumplir es la transformación trivial $t(\underline{x}) = \underline{x}$, sin embargo el objetivo de plantear este supuesto es encontrar otra transformación distinta a la trivial que permita que los siguientes dos supuestos se cumplan. Este supuesto implica la equivalencia entre encontrar la distribución $p(\underline{x}|G_i, D)$ y la distribución $p(t(\underline{x})|G_i, D)$, por lo cual a partir de este momento el análisis se concentra en encontrar esta segunda distribución.

B) Dentro de cada grupo las transformadas de los individuos pueden modelarse como si fueran intercambiables, lo cual quiere decir que la función de distribución conjunta de

las transformadas $\{\underline{t}_{i1}, \underline{t}_{i2}, \ldots, \underline{t}_{im} : \underline{t}_{ij} = t(\underline{x}_{ij})\}$ de los individuos del grupo i es invariante ante permutaciones. La utilidad de este supuesto radica en que, cuando se cumple, por medio del teorema de representación de DeFinetti la función de distribución conjunta de las transformadas $\{\underline{t}_{i1}, \underline{t}_{i2}, \ldots, \underline{t}_{im} : \underline{t}_{ij} = t(\underline{x}_{ij})\}$ para el grupo i se puede expresar como:

$$p(\underline{t}_{i1}, \underline{t}_{i2}, \ldots, \underline{t}_{im}|G_i) = \int_{\Theta} p(\underline{\theta}_i|G_i) \prod_{j=1}^{m} p(\underline{t}_{ij}|\underline{\theta}_i, G_i) d\underline{\theta}_i$$

$$= \int_{\Theta} p(\underline{\theta}_i|G_i) \prod_{j=1}^{m} p(\underline{t}_{ij}|\underline{\theta}_i) d\underline{\theta}_i$$

Es claro que de esta expresión se puede deducir que la distribución $p(t(\underline{x})|G_i, D)$ para un nuevo individuo se puede expresar entonces como:

$$p(t(\underline{x})|G_i, D) = \int_{\Theta} p(\underline{\theta}_i|G_i, D) p(t(\underline{x})|\underline{\theta}_i) d\underline{\theta}_i \ ,$$

donde por el teorema de Bayes:

$$p(\underline{\theta}_i|G_i, D) \propto \prod_{j=1}^{n_i} p(\underline{t}_{ij}|\underline{\theta}_i) p(\underline{\theta}_i|G_i) \qquad \text{siendo } n_i \text{ la cardinalidad del grupo i-ésimo .}$$

Resta entonces encontrar la especificación de $p(\underline{\theta}_i|G_i)$ y la de $p(\underline{t}_{ij}|\underline{\theta}_i)$ que se utilizará también para $p(t(\underline{x})|\underline{\theta}_i)$.

C) El tercer supuesto consiste en que la transformación $t = t(\underline{x})$ se seleccione de forma que tanto $p(\underline{t}_{ij}|\underline{\theta}_i)$ como $p(t(\underline{x})|\underline{\theta}_i)$ asuman una distribución Normal l-variada bajo el siguiente esquema:

$$p(t(\underline{x})|\underline{\theta}_i) = N_l(t(\underline{x})|M_i, H_i) \ .$$

Además, se plantea el uso de una distribución a priori no informativa de referencia que corresponde a:

$$p(\underline{\theta}_i|G_i) = \Pi(M_i, H_i) \propto |H_i|^{\frac{-\nu}{2}} \ \ (\nu \in R^+ \textit{conocido}) \ .$$

En resumen, por la expresión (1) y los supuestos A) y B) se tiene que:

$$p(G_i|\underline{x}, D) \approx p(G_i|t(\underline{x}), D) \propto p(G_i|D) p(t(\underline{x})|G_i, D) \text{ para cada } i \in J_k \qquad (2)$$

donde

$$p(t(\underline{x})|G_i, D) = \int_{\Theta} p(\underline{\theta}_i|G_i, D) p(t(\underline{x})|\underline{\theta}_i, D) d\underline{\theta}_i$$

y

$$p(\underline{\theta}_i|G_i, D) \propto \prod_{j=1}^{n_i} p(\underline{t}_{ij}|\underline{\theta}_i)p(\underline{\theta}_i|G_i) \ .$$

Por el supuesto C) se tiene:

$$p(G_i|t(\underline{x}), D) \propto p(G_i|D) \int\int \Pi(\underline{\theta}_i|G_i)[\prod_{j=1}^{n_i} N_l(t_{ij}|M_i, H_i)]N_l(t(\underline{x})|M_i, H_i)dM_i dH_i \ .$$

Por medio de desarrollos algebraicos expresados en el apéndice I se tiene que:

$$p(G_i|t(\underline{x}), D) \propto p(G_i|D)St_l(t(\underline{x})|n_i - \nu + 1, \bar{\underline{t}}_i, \frac{n_i - \nu + 1}{n_i + 1}S_i^{-1}) \tag{3}$$

donde

$$\bar{\underline{t}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \underline{t}_{ij}$$

y

$$S_i = \frac{1}{n_i}\sum_{j=1}^{n_i} (\underline{t}_{ij} - \bar{\underline{t}}_i)(\underline{t}_{ij} - \bar{\underline{t}}_i)'$$

son el vector de medias y la matriz de varianzas-covarianzas de las transformadas del grupo i respectivamente.

La transformación que se propone utilizar es tal que $t : R^r \to R^{k-1}$ y se define como:

$$t(\underline{x}) = (\lambda_1(\underline{x}), \lambda_2(\underline{x}), \ldots, \lambda_{k-1}(\underline{x}))$$

en donde:

$$\lambda_i(\underline{x}) = a_i'\underline{x} = (\overline{\underline{x}}_i - \overline{\underline{x}}_{(i)})'V^{-1}\underline{x} \text{ para cada } i \in J_k \tag{4}$$

con

$$\overline{\underline{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \underline{x}_{ij},$$

$$\overline{\underline{x}}_{(i)} = \frac{1}{n - n_i}\sum_{l \neq i}\sum_{j=1}^{n_l} \underline{x}_{lj}$$

y

$$V = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} (\underline{x}_{ij} - \overline{\underline{x}}_i)(\underline{x}_{ij} - \overline{\underline{x}}_i)' \ .$$

La razón para proponer esta transformación se basa en la similitud con la idea planteada por Fisher de buscar maximizar las distancias entre grupos estandarizadas por la varianza de los mismos bajo el supuesto de igualdad de varianzas. Aquí se busca maximizar las distancias entre los individuos de un grupo y los que no pertenecen a dicho grupo.

Una vez obtenidos los valores de las transformaciones, y utilizadas éstas para obtener las probabilidades predictivas, el individuo nuevo será asignado al grupo para el que la correspondiente probabilidad predictiva sea máxima (Bernardo, 1988).

Capítulo IV. Aplicación de las Técnicas de solución al problema de agrupación a la clasificación de la población por su condición de pobreza

a. Selección de los indicadores

En el contexto del problema de clasificación de los hogares por su condición de pobreza, el consenso general entre los estudiosos de la materia en el ámbito económico ha señalado que además de una falta de ingresos, existen diversos aspectos de influencia a dicha condición, como lo son:

a) Inadecuada alimentación (desnutrición, riesgos de salud crónicos)

b) Malas condiciones de vida (malas viviendas y falta de acceso a servicios básicos)

c) Bajas expectativas de mejora a futuro (poco o nulo acceso a educación y salud que permitan mejorar la condición de vida)

d) Otros elementos que determinan la pobreza (inseguridad, discriminación, falta de acceso a mercados laborales)

Por ello y con el objetivo de aplicar el método de enfoque multidimensional de medición de la pobreza, se plantea la necesidad de seleccionar algunos indicadores que servirán para identificar a los hogares en condición de pobreza, de entre todos los indicadores que se encuentran disponibles, y que se relacionen con los cuatro aspectos previamente señalados.

Dado que se trabajará con una encuesta que no está específicamente diseñada para medir los aspectos que se definieron previamente, para algunos de ellos no se encontrarán indicadores que los representen, lo cual conlleva a la posibilidad de medir de manera incompleta el concepto de pobreza.

Debido a que el ingreso es considerado como una medida eficiente de la pobreza cuando es confiable su captación, como es el caso de la ENIGH, se comparó el resto de los indicadores con el ingreso de los hogares para evaluar su utilidad en la obtención de una clasificación de la población por su condición de pobreza. Se realizó un análisis de correspondencias para cada una de las variables categóricas con respecto a los deciles de

ingreso, con el objetivo de identificar si se puede reducir el número de categorías de algunas variables. Posteriormente para evaluar qué variables proporcionan mayor información para identificar a los hogares en condiciones de pobreza, se obtuvieron las medidas de asociación entre las variables categóricas y dichos deciles, así como las regresiones entre el ingreso y las variables continuas, cuantificando su poder de clasificación.

Estos pasos permitirán establecer un conjunto de indicadores útiles para el análisis multidimensional que reflejen de manera similar al ingreso las condiciones de vida, bajo la premisa de que el ingreso es un buen indicador de dicha condición de vida.

i) Medidas de Asociación

En el caso de la ENIGH se encontró que en todos los casos se rechaza la hipótesis de independencia entre las variables categóricas elegidas y los deciles de ingreso.

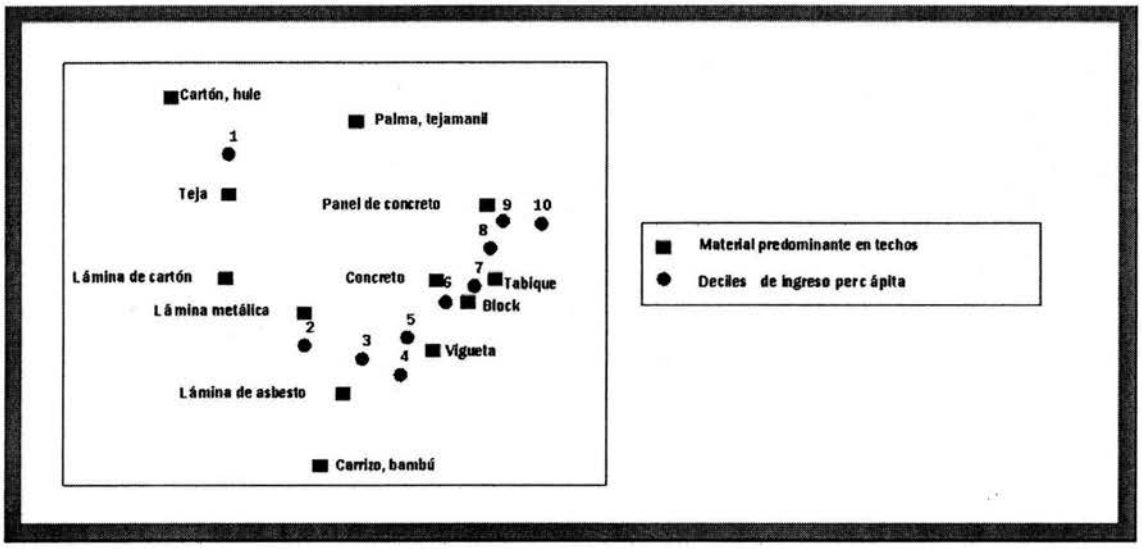
En la tabla siguiente se presenta el coeficiente de contingencia en media cuadrática de Pearson de cada una de las variables categóricas con respecto a los deciles de ingreso en orden descendente.

Se observa que las variables más asociadas con el ingreso son las características de las viviendas y algunos enseres mayores, mientras que las menos asociadas son los enseres menores. Esta tabla proporciona uno de los elementos para decidir cuáles variables serán tomadas en cuenta en los análisis subsecuentes.

Coeficiente de Contingencia en media cuadrática de Pearson entre la variable referida y los deciles de ingreso per cápita	
Variable	Coeficiente
Material del piso	0.530
Combustible que utiliza	0.524
Disponibilidad de agua	0.506
Tamaño de la localidad de residencia	0.482
Drenaje	0.473
Servicio Sanitario	0.472
Refrigerador	0.469
Teléfono	0.466
Calentador de gas	0.466
Estufa de gas	0.462
Vehículo	0.441
Recolección pública de basura	0.440
Material de los techos	0.428
Computadora	0.427
Licuadora	0.420
Videocasetera	0.418
Lavadora	0.412
Posesión de la vivienda	0.406
Material de los muros	0.399
Tocadiscos	0.382
Acceso a seguridad social	0.382
Televisión	0.379
Jefatura analfabeta	0.307
Ventilador	0.292
Radiograbadora	0.235
Luz	0.232
Radio	0.114
Estufa de otro combustible	0.110
Calentador de otro combustible	0.086
Grabadora	0.083
Jefatura femenina	0.078

ii) Análisis de Correspondencias Simples

A continuación se presenta uno de los análisis de correspondencias que se realizaron:



En este análisis se observa que categorías como tabique, block y concreto podrían ser fusionadas en una sola categoría dado que reflejan prácticamente la misma condición con respecto al ingreso. En el Apéndice II se incluyen todos los análisis de correspondencias simples que se obtuvieron, y que sirvieron como base para la reducción del número de categorías de algunas de las variables además de reflejar la asociación de manera visual (con mayor claridad) entre las variables categóricas y los deciles de ingreso.

iii) Regresión para variables continuas

Para las variables continuas se llevaron a cabo análisis de regresión, pero no contra el ingreso per cápita sino contra el logaritmo natural del ingreso per cápita dado que con esta transformación se logra que los supuestos necesarios para la regresión sean válidos. Los resultados de los análisis de regresión lineal simple se reflejan en la siguiente tabla que contiene la estadística R cuadrada, misma que analizada en términos comparativos muestra cuáles son las variables de mayor relación con el ingreso.

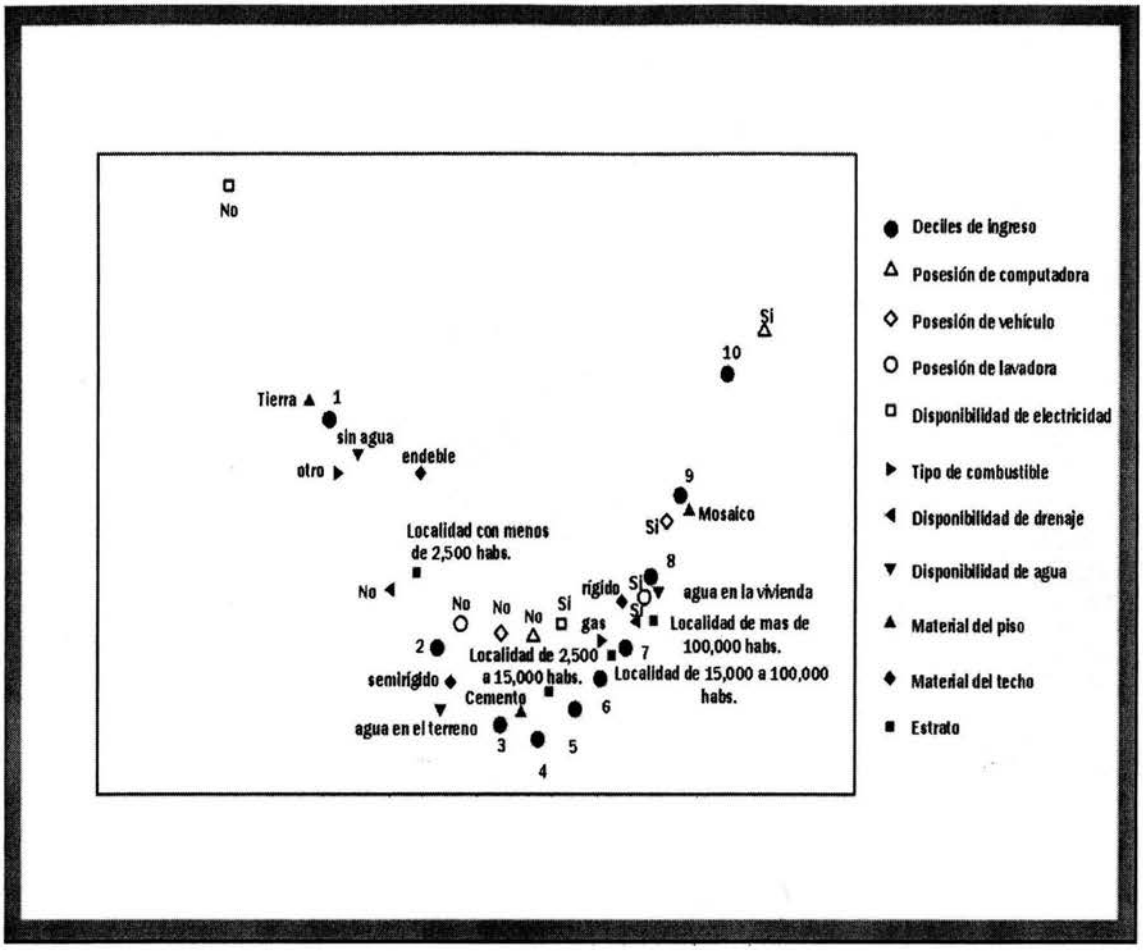
Estadística R cuadrada para la regresión con respecto al logaritmo natural del ingreso mensual per cápita de algunas variables de interés	
Variable	R cuadrada
Años de escolaridad del jefe del hogar	0.292
Índice de hacinamiento (personas por cuarto)	0.269
Número de cuartos	0.185
Menores de 15 años	0.173
Niños de 5 a 15 años en el hogar	0.147
Índice de dependencia demográfica (personas no en edad económicamente activa entre personas en edad económicamente activa)	0.137
Total de personas	0.121
Índice de dependencia (personas que no trabajan por cada trabajador)	0.105
Total de mujeres	0.075
Total de hombres	0.069
Mayores de 65 años	0.006
Edad del jefe del hogar	0.001

Con la información de cuadro previo, se decide tomar a los indicadores compuestos para realizar los análisis, y por ello para evitar colinearidad, se descartan las variables número de cuartos, personas en el hogar (por edades, por sexo y total) así como el índice de dependencia no demográfico. Un análisis de regresión múltiple para el resto de las variables ratifica su utilidad predictiva, resultando todas significativas.

iv) Análisis de correspondencias múltiples

Con los análisis previos se comienzan a distinguir las variables que pueden ser útiles en la clasificación de los hogares por su condición de pobreza. Ahora bien, es muy importante recordar que si bien una variable a nivel individual puede tener una asociación alta con respecto a otra, cuando se incorporan al análisis otras variables, estos niveles de asociación pueden reducirse. Por ello es muy importante que se analicen las variables en su conjunto, para poder determinar así cuáles de ellas serán las que se tomen en cuenta en los análisis posteriores.

A continuación se presenta la gráfica del análisis de correspondencias múltiples para algunas de las variables de mayor asociación con los deciles de ingreso mensual per cápita.



Se observa un claro ordenamiento generado principalmente por la distribución del ingreso, donde a partir de la ubicación de los deciles se puede determinar una relación cuadrática. A partir de dicha relación se analizan el resto de los indicadores, teniendo marcadas relaciones de orden entre las características del piso de la vivienda y los deciles por ejemplo. Otro punto de interés es observar cuales indicadores son básicos, esto es, cuáles son los que se adquieren primero, y cuáles son de lujo, es decir, los que se adquieren cuando ya se ha alcanzado un nivel de ingresos mayor. En este renglón se observa que la luz es uno de los primeros servicios que se adquieren cuando no se dispone de ella, mientras que la computadora es de los últimos enseres adquiridos.

De los resultados previos se decide elegir al siguiente conjunto de indicadores para realizar los análisis subsecuentes: años de escolaridad del jefe del hogar, índice de hacinamiento (promedio de personas por cuarto en el hogar), índice de dependencia demográfica (personas que no están en edad económicamente activa entre personas en edad económicamente activa), tamaño de la localidad de residencia, material de los techos (codificado en endebles, semirígidos y rígidos), material de los pisos (codificado en de tierra,

de cemento y de otro recubrimiento), agua (codificado en sin agua, agua dentro del terreno pero fuera de la vivienda y agua dentro de la vivienda), disponibilidad de drenaje, uso de gas como combustible, disponibilidad de electricidad, posesión de lavadora, posesión de vehículo y posesión de computadora.

b. Construcción de una clasificación inicial

Como se mencionó en el capítulo 3 se construyen dos clasificaciones iniciales, la primera obtenida directamente de la aplicación de la propuesta del Comité Técnico de Medición de la pobreza, que genera los grupos siguientes:

Clasificación inicial de acuerdo con la propuesta del Comité Técnico de Medición de la Pobreza	
Grupo	**Porcentaje**
Grupo 1 (Personas sin ingreso suficiente para alimentación adecuada)	28.9%
Grupo 2 (Personas sin ingreso suficiente para educación y salud adecuados)	7.6%
Grupo 3 (Personas sin ingreso suficiente para vivienda, vestido, calzado, transporte)	21.8%
Grupo 4 (Personas con ingreso suficiente para cubrir sus necesidades)	41.7%

La segunda de las clasificaciones iniciales requerida para aplicar las técnicas estadísticas de agrupación se obtuvo utilizando el algoritmo de "simulated annealing" a una agrupación inicial que en este caso se definió a partir de los centiles del ingreso per cápita de los hogares.

La razón para elegir los centiles de ingreso está sustentada en la teoría económica que en primera instancia señala al consumo como el mejor indicador del bienestar de un hogar, y al ingreso como su mejor aproximación, esto es para ser consistentes con los objetivos del trabajo.

La función de utilidad que se usa para aplicar el algoritmo es: p*log(p) con p siendo la probabilidad de que un individuo sea asignado a un grupo determinado. Para implementar el algoritmo se corrió un análisis de componentes principales para las variables involucradas, y se eligieron los 14 componentes (recordar que algunas de las variables generan dummies) que más varianza explican (92.38 por ciento).

En el apéndice III se presenta el programa que se utilizó para implementar el procedimiento. Los resultados obtenidos se pueden apreciar en la siguiente distribución porcentual de los grupos:

Clasificación inicial de acuerdo con los resultados del simulated annealing	
Grupo	Porcentaje
Grupo 1	7.0%
Grupo 2	1.0%
Grupo 3	92.0%

Esta distribución como puede apreciarse genera tres grupos de los cuales dos corresponden al porcentaje de la población con ingresos más bajos. Esto lleva a pensar que los indicadores elegidos son útiles para distinguir entre los hogares que definitivamente tienen carencias extraordinarias con respecto del resto de la población. Si se rescata la clasificación para cuatro grupos de iteraciones previas, se obtiene la siguente distribución porcentual de los grupos:

Clasificación inicial de acuerdo con los resultados del simulated annealing (cuatro grupos)	
Grupo	Porcentaje
Grupo 1	7.0%
Grupo 2	1.0%
Grupo 3	91.0%
Grupo 4	1.0%

Esta distribución muestra que las variables indicadas también son de utilidad para distinguir a los hogares con mayores ingresos. Sin embargo el tercer grupo sigue apareciendo muy grande y todavía sin distinciones, por lo cual se continúa recuperando iteraciones previas para intentar obtener distinciones del grupo 3 que sean de utilidad, así, si se toma la separación en cinco grupos se obtiene la siguiente distribución porcentual de los grupos:

Clasificación inicial de acuerdo con los resultados del simulated annealing (cinco grupos)	
Grupo	Porcentaje
Grupo 1	7.0%
Grupo 2	1.0%
Grupo 3	32.0%
Grupo 4	59.0%
Grupo 5	1.0%

Estos resultados muestran que los datos primero sugieren una distinción de los hogares con mayores carencias, después la distinción entre los hogares con mayores riquezas, y una vez dadas estas dos separaciones, se empieza a generar la clasificación del resto de los individuos. En el caso de la separación en seis grupos esta sería la distribución porcentual de los grupos:

Clasificación inicial de acuerdo con los resultados del simulated annealing (seis grupos)	
Grupo	Porcentaje
Grupo 1	7.0%
Grupo 2	1.0%
Grupo 3	32.0%
Grupo 4	27.0%
Grupo 5	32.0%
Grupo 6	1.0%

La existencia de dos grupos de tamaño uno por ciento refleja la existencia de valores atípicos en dichos grupos o en los grupos aledaños (dado que no permitieron que se unieran a ellos). Es lógico que existan dichos valores tanto en el primer como en el último grupo dado que corresponden a los hogares extremadamente pobres o ricos respectivamente. Por ello, y dado que para efectos prácticos la presencia de dos grupos con uno por ciento de la población no es informativo, los grupos que tienen dicho porcentaje serán asimilados al grupo aledaño, por lo que se trabajará para esta propuesta con la siguiente distribución porcentual de los grupos:

Clasificación inicial de acuerdo con los resultados del simulated annealing (seis grupos asimilados a cuatro)	
Grupo	Porcentaje
Grupo 1	8.0%
Grupo 2	32.0%
Grupo 3	27.0%
Grupo 4	33.0%

La interpretación de los grupos del cuadro anterior consiste en que los datos por sí solos generan una agrupación óptima que se refleja en dichos porcentajes, dado que los grupos están ordenados por ingreso, se plantea tentativamente la existencia de un grupo de hogares en pobreza extrema, un grupo en pobreza moderada, un grupo con condición media de vida y un grupo con características satisfactorias. Es importante resaltar que esta es una definición de pobreza que los datos están sugiriendo. Es clara la diferencia con la propuesta del Comité, sin embargo un punto a destacar es que una reagrupación juntando los grupos uno y dos por una parte, y los grupos tres y cuatro por la otra generaría resultados similares entre las dos propuestas, del mismo modo que el agrupar los primeros tres grupos y dejar libre al cuarto genera resultados semejantes.

Lo anterior suena lógico si se recuerda que el primer corte de la agrupación del Comité se refiere a cuestiones relacionadas con la alimentación, lo cual no está medido en los datos incorporados en el procedimiento de "simulated annealing", pero el segundo y tercer cortes de la agrupación del Comité involucra cuestiones relacionadas con educación y aspectos relacionados con la vivienda, lo cual sí está incorporado en los datos multivariados.

c. Aplicación de las Técnicas utilizando la primera alternativa de clasificación inicial

Para cumplir el objetivo de evaluar la propuesta de clasificación de la población por su condición de pobreza hecha por el Comité Técnico para la Medición de la Pobreza se aplicaron, utilizando dicha propuesta como la primera alternativa de clasificación inicial, las técnicas de análisis discriminante, regresión logística y la metodología bayesiana descritas en el capítulo anterior. En el apéndice IV se muestra el programa que implementa la metodología bayesiana.

Las tres técnicas muestran un porcentaje de clasificación similar de alrededor de un 60 por ciento, lo cual muestra que al intentar clasificar a un hogar por medio de las

variables que en este trabajo resultaron importantes para definir su condición de pobreza se estaría incurriendo en una inconsistencia del 40 por ciento de los casos con respecto a la clasificación por ingreso.

Porcentaje de clasificación total correcta por método utilizado	
Método	**Porcentaje**
Análisis Discriminante	63.9%
Regresión logística	60.0%
Metodología Bayesiana	63.1%

Este resultado se puede interpretar desde dos perspectivas, por una parte puede estar ocurriendo que la clasificación por ingreso no está reflejando lo mismo que lo que los datos multivariados muestran, y por otra parte puede estar ocurriendo que los datos no sean suficientemente explicativos de la clasificación por ingreso ya sea por falta de información o por distorsiones normales debidas a la libertad de elección de los individuos con respecto a ejercer su posibilidad de reflejar un mejor ingreso en su condición de vida.

Intentando determinar cuál de las perspectivas es más viable, se analiza por una parte el porcentaje de hogares resultante en cada grupo dependiendo de la técnica utilizada, y por otra parte, se observa si la variación porcentual de aquéllos hogares que cambian de grupo de acuerdo a los datos (y a la técnica) es radical o es sutil.

Distribución de la clasificación por grupos según metodología				
Grupo	**Original**	**Análisis Discriminante**	**Regresión Logística**	**Metodología Bayesiana**
1	28.9%	20.6%	22.4%	19.8%
2	7.6%	0.0%	0.0%	0.0%
3	21.8%	14.8%	11.0%	4.9%
4	41.7%	64.6%	66.7%	75.3%

En la tabla anterior se puede apreciar que la distribución porcentual de los hogares clasificados en los distintos grupos es sensible a la técnica estadística que se utiliza, pero que en cualquier caso existe una tendencia a polarizar los grupos, desapareciendo incluso el segundo que desde un principio tiene un tamaño reducido. Ahora se observará hacia donde se mueven los hogares que cambian de clasificación.

Cuadro de porcentaje de clasificación original contra predicho (Análisis Discriminante)					
Grupo Original	Predicho				
	1	2	3	4	Total
1	62.8%	0.0%	19.6%	17.6%	100.0%
2	34.0%	0.0%	26.5%	39.5%	100.0%
3	16.0%	0.0%	24.9%	59.1%	100.0%
4	4.2%	0.0%	7.2%	88.5%	100.0%

Cuadro de porcentaje de clasificación original contra predicho (Regresión Logística)					
Grupo Original	Predicho				
	1	2	3	4	Total
1	67.0%	0.0%	13.8%	19.3%	100.0%
2	37.0%	0.0%	20.0%	43.0%	100.0%
3	17.9%	0.0%	19.6%	62.4%	100.0%
4	4.8%	0.0%	5.2%	90.0%	100.0%

Cuadro de porcentaje de clasificación original contra predicho (Metodología Bayesiana)					
Grupo Original	Predicho				
	1	2	3	4	Total
1	61.6%	0.0%	7.6%	30.8%	100.0%
2	32.3%	0.0%	7.7%	60.1%	100.0%
3	14.5%	0.0%	8.7%	76.7%	100.0%
4	4.0%	0.0%	2.0%	94.0%	100.0%

Existe una consistencia entre técnicas en lo que concierne a los cambios de clasificación, observándose que tanto los hogares del primero como los del cuarto grupo conservan en la mayoría de los casos su clasificación. En el caso del segundo grupo que desaparece, los hogares se reclasifican en mayor medida dentro del cuarto grupo, finalmente para el tercer grupo se aprecia un alto porcentaje de reclasificación reagrupándose en su mayoría en el cuarto grupo.

Para realizar una mejor evaluación de la propuesta del Comité se realizó nuevamente el análisis previo utilizando clasificaciones de dos grupos, esto es, planteando al primer grupo y a la unión de los restantes como un par, después juntando los primeros dos grupos y juntando los últimos dos grupos, y finalmente juntando los primeros tres grupos y dejando

libre al último. Esto se realiza con dos objetivos, el primero es evaluar si la clasificación con menos grupos es más robusta, y segundo porque en la práctica el análisis de la pobreza puede llevarse a cabo desde la perspectiva de comparar los pobres por cada concepto de manera separada.

Los resultados que se presentan a continuación muestran que para la primera agrupación (primera línea de pobreza), se tiene un muy alto porcentaje de clasificación correcto global, sin embargo desequilibrado dado que el grupo que corresponde a la no pobreza se clasifica muy bien, mientras que el que corresponde a la pobreza tiene en todas las técnicas un porcentaje de clasificación incorrecta de aproximadamente la mitad de los casos.

Porcentaje de clasificación total correcta por método utilizado

Método	Porcentaje
Análisis Discriminante	86.0%
Regresión logística	86.1%
Metodología Bayesiana	85.9%

Distribución de la clasificación por grupos según metodología

Grupo	Original	Análisis Discriminante	Regresión Logística	Metodología Bayesiana
1	28.9%	15.8%	14.3%	12.3%
2, 3 y 4	71.1%	84.2%	85.7%	87.7%

Cuadro de porcentaje de clasificación original contra predicho (Análisis Discriminante)

Grupo Original	Predicho		
	1	2, 3 y 4	Total
1	54.5%	45.5%	100.0%
2, 3 y 4	6.0%	94.0%	100.0%

Cuadro de porcentaje de clasificación original contra predicho (Regresión logística)

Grupo Original	Predicho		
	1	2, 3 y 4	Total
1	51.0%	49.0%	100.0%
2, 3 y 4	5.0%	95.0%	100.0%

Cuadro de porcentaje de clasificación original contra predicho (Metodología Bayesiana)			
Grupo Original	Predicho		
	1	2,3 y 4	Total
1	45.4%	54.6%	100.0%
2,3 y 4	3.8%	96.2%	100.0%

En el caso de la segunda agrupación (segunda línea de pobreza) con porcentaje de clasificación menor al anterior, el porcentaje de clasificación del grupo correspondiente a la no pobreza se conserva alto mientras que el del grupo correspondiente a la pobreza mejora ligeramente.

Porcentaje de clasificación total correcta por método utilizado	
Método	Porcentaje
Análisis Discriminante	82.7%
Regresión logística	83.0%
Metodología Bayesiana	82.5%

Distribución de la clasificación por grupos según metodología				
Grupo	Original	Análisis Discriminante	Regresión Logística	Metodología Bayesiana
1 y 2	36.5%	20.7%	21.2%	17.0%
3 y 4	63.5%	79.3%	78.8%	83.0%

Cuadro de porcentaje de clasificación original contra predicho (Análisis Discriminante)			
Grupo Original	Predicho		
	1 y 2	3 y 4	Total
1 y 2	56.3%	43.7%	100.0%
3 y 4	7.4%	92.6%	100.0%

Cuadro de porcentaje de clasificación original contra predicho (Regresión logística)			
Grupo Original	Predicho		
	1 y 2	3 y 4	Total
1 y 2	57.8%	42.2%	100.0%
3 y 4	7.6%	92.4%	100.0%

Cuadro de porcentaje de clasificación original contra predicho (Metodología Bayesiana)			
Grupo Original	Predicho		
	1 y 2	3 y 4	Total
1 y 2	49.2%	50.8%	100.0%
3 y 4	5.1%	94.9%	100.0%

Para la tercera agrupación, con un porcentaje global de clasificación aún menor, se observa sin embargo un buen porcentaje de clasificación en ambos grupos.

Porcentaje de clasificación total correcta por método utilizado	
Método	Porcentaje
Análisis Discriminante	76.7%
Regresión logística	77.7%
Metodología Bayesiana	76.7%

Distribución de la clasificación por grupos según metodología				
Grupo	Original	Análisis Discriminante	Regresión Logística	Metodología Bayesiana
1, 2 y 3	58.3%	45.2%	45.8%	43.3%
4	41.7%	54.8%	54.2%	56.7%

Cuadro de porcentaje de clasificación original contra predicho (Análisis Discriminante)			
Grupo Original	Predicho		
	1, 2 y 3	4	Total
1, 2 y 3	72.7%	27.3%	100.0%
4	19.7%	80.3%	100.0%

Cuadro de porcentaje de clasificación original contra predicho (Regresión logística)			
Grupo Original	Predicho		
	1, 2 y 3	4	Total
1, 2 y 3	74.2%	25.8%	100.0%
4	19.5%	80.5%	100.0%

Cuadro de porcentaje de clasificación original contra predicho (Metodología Bayesiana)			
Grupo Original	Predicho		
	1, 2 y 3	4	Total
1, 2 y 3	70.8%	29.2%	100.0%
4	17.8%	82.2%	100.0%

En todos los casos se observa que la distribución porcentual de la clasificación resultante se distorsiona con respecto a la distribución porcentual original.

Se observa que al realizar el análisis de manera separada, la primera agrupación proporciona los mayores porcentajes de clasificación, lo que concuerda con el análisis conjunto para los cuatro grupos, sin embargo, la clasificación mas equilibrada parece ser la que junta los primeros tres grupos. En este punto se puede apreciar una complicación más del problema, al tener que tomar la decisión sobre el interés de tener una clasificación conjunta óptima o una clasificación óptima grupo a grupo.

Se analizaron algunas alternativas a la propuesta del Comité que partieron de su diseño conceptual e incluso de la propuesta en sí, como la agrupación por quintiles o deciles de ingreso, o el intento de separar grupos de la propuesta, sin embargo ninguno de los intentos generó una clasificación que se pudiera considerar que mejoraba la propuesta del Comité, por lo tanto fueron descartadas.

d. Aplicación de las Técnicas utilizando la segunda alternativa de clasificación inicial

En este apartado se replica el análisis del apartado anterior pero utilizando la segunda alternativa de clasificación. Los resultados se presentan bajo la misma estructura. Se aplicaron las técnicas de análisis discriminante, regresión logística y la metodología bayesiana.

Las tres técnicas muestran un porcentaje de clasificación similar de entre un 55 y un 60 por ciento, ligeramente menor al correspondiente a la primera alternativa.

Porcentaje de clasificación total correcta por método utilizado	
Método	**Porcentaje**
Análisis Discriminante	56.7%
Regresión logística	59.3%
Metodología Bayesiana	55.7%

Una vez más se analizan por una parte el porcentaje de hogares resultante en cada grupo dependiendo de la técnica utilizada, y por otra parte, la variación porcentual de aquéllos hogares que cambian de grupo de acuerdo a los datos (y a la técnica).

Distribución de la clasificación por grupos según metodología				
Grupo	Original	Análisis Discriminante	Regresión Logística	Metodología Bayesiana
1	8.0%	9.2%	4.5%	3.4%
2	32.0%	31.7%	37.5%	38.3%
3	27.0%	25.0%	20.2%	17.1%
4	33.0%	34.2%	37.8%	41.2%

En este caso se aprecia que ningún grupo desaparece, y que la distribución porcentual de los hogares clasificados en los distintos grupos si bien es sensible a la técnica estadística que se utiliza, conserva en cualquier caso las proporciones originales con cambios pequeños, sustancialmente menores a los encontrados en el análisis de la primera alternativa, principalmente en la técnica del análisis discriminante. A continuación se muestra hacia donde se mueven los hogares que cambian de clasificación.

Cuadro de porcentaje de clasificación original contra predicho (Análisis Discriminante)					
Grupo Original	Predicho				
	1	2	3	4	Total
1	58.6%	36.9%	3.4%	1.1%	100.0%
2	12.7%	57.9%	22.1%	7.4%	100.0%
3	1.2%	29.6%	37.9%	31.3%	100.0%
4	0.2%	6.6%	22.5%	70.7%	100.0%

Cuadro de porcentaje de clasificación original contra predicho (Regresión Logística)					
Grupo Original	Predicho				
	1	2	3	4	Total
1	36.5%	58.6%	3.9%	1.0%	100.0%
2	4.6%	68.2%	19.5%	7.7%	100.0%
3	0.2%	31.9%	32.6%	35.3%	100.0%
4	0.0%	7.2%	14.7%	78.1%	100.0%

Cuadro de porcentaje de clasificación original contra predicho (Metodología Bayesiana)					
Grupo Original	Predicho				
	1	2	3	4	Total
1	30.0%	65.6%	3.0%	1.4%	100.0%
2	3.1%	66.5%	18.1%	12.4%	100.0%
3	0.1%	33.1%	24.0%	42.9%	100.0%
4	0.0%	8.5%	13.9%	77.6%	100.0%

En todas las técnicas se observa que el cuarto grupo se clasifica correctamente en un alto porcentaje, sin embargo para el resto de los grupos la clasificación se intercambia notoriamente, casi nunca de manera radical, sino más bien en las fronteras. Estos cambios son más sutiles en el análisis discriminante, pero no por ello menos importantes. El tercer grupo se confunde con el segundo y el cuarto, el segundo grupo se clasifica mejor, pero el primer grupo se confunde con el segundo. En un sentido práctico, el que los errores no sean radicales no es tan problemático, para fines de interpretación esto significa que los hogares de las fronteras pueden intercambiarse de grupo con mayor facilidad.

Replicando los análisis del apartado anterior, en los que se utilizan clasificaciones de dos grupos, planteando al primer grupo y a la unión de los restantes como un par, después juntando los primeros dos grupos y juntando los últimos dos grupos, y finalmente juntando los primeros tres grupos y dejando libre al último, se observa que para la primera agrupación (primera línea de pobreza), se tiene un porcentaje de clasificación correcto global aún mayor al de la primera alternativa, sin embargo conservando el carácter de desequilibrio puesto que el grupo que corresponde a la no pobreza sigue clasificándose muy bien, mientras que el que corresponde a la pobreza sigue teniendo en todas las técnicas un porcentaje de clasificación incorrecta de menos del 60 por ciento de los casos. Cabe señalar que la distribución porcentual original se conserva en mayor medida en comparación con la de la primera alternativa.

Porcentaje de clasificación total correcta por método utilizado	
Método	Porcentaje
Análisis Discriminante	92.5%
Regresión logística	93.4%
Metodología Bayesiana	93.2%

Distribución de la clasificación por grupos según metodología				
Grupo	Original	Análisis Discriminante	Regresión Logística	Metodología Bayesiana
1	8.0%	8.4%	4.2%	5.4%
2, 3 y 4	92.0%	91.6%	95.8%	94.6%

Cuadro de porcentaje de clasificación original contra predicho (Análisis Discriminante)			
Grupo Original	Predicho		
	1	2, 3 y 4	Total
1	56.2%	43.8%	100.0%
2, 3 y 4	4.3%	95.7%	100.0%

Cuadro de porcentaje de clasificación original contra predicho (Regresión logística)			
Grupo Original	Predicho		
	1	2, 3 y 4	Total
1	34.9%	65.1%	100.0%
2, 3 y 4	1.5%	98.5%	100.0%

Cuadro de porcentaje de clasificación original contra predicho (Metodología Bayesiana)			
Grupo Original	Predicho		
	1	2,3 y 4	Total
1	41.1%	58.9%	100.0%
2,3 y 4	2.3%	97.7%	100.0%

De manera similar al caso de la primera agrupación, en el caso de la segunda (segunda línea de pobreza) se tiene una distribución porcentual en la clasificación generada por las técnicas muy similar a la original (lo cual no ocurre en la primera alternativa de clasificación inicial). Además, el porcentaje de clasificación del grupo correspondiente a la no pobreza

se conserva alto, y el del grupo correspondiente a la pobreza mejora en mayor medida que lo observado en la primera alternativa.

Porcentaje de clasificación total correcta por método utilizado	
Método	**Porcentaje**
Análisis Discriminante	80.0%
Regresión logística	80.3%
Metodología Bayesiana	79.6%

Distribución de la clasificación por grupos según metodología				
Grupo	**Original**	**Análisis Discriminante**	**Regresión Logística**	**Metodología Bayesiana**
1 y 2	40.0%	35.3%	36.2%	32.2%
3 y 4	60.0%	64.7%	63.8%	67.8%

Cuadro de porcentaje de clasificación original contra predicho (Análisis Discriminante)			
Grupo Original	**Predicho**		
	1 y 2	**3 y 4**	**Total**
1 y 2	69.2%	30.8%	100.0%
3 y 4	12.8%	87.2%	100.0%

Cuadro de porcentaje de clasificación original contra predicho (Regresión logística)			
Grupo Original	**Predicho**		
	1 y 2	**3 y 4**	**Total**
1 y 2	70.6%	29.4%	100.0%
3 y 4	13.3%	86.7%	100.0%

Cuadro de porcentaje de clasificación original contra predicho (Metodología Bayesiana)			
Grupo Original	**Predicho**		
	1 y 2	**3 y 4**	**Total**
1 y 2	64.9%	35.1%	100.0%
3 y 4	10.5%	89.5%	100.0%

Para la tercera agrupación, el porcentaje global de clasificación se conserva alto, aún

más que en la primera alternativa. La distribución porcentual se conserva en todas las técnicas, lo que no ocurre en la primera alternativa, sin embargo es muy interesante observar que en este caso el porcentaje de clasificación correcto para el primer grupo es alto mientras que el del segundo grupo ya no lo es tanto.

Porcentaje de clasificación total correcta por método utilizado

Método	Porcentaje
Análisis Discriminante	80.5%
Regresión logística	82.0%
Metodología Bayesiana	80.4%

Distribución de la clasificación por grupos según metodología

Grupo	Original	Análisis Discriminante	Regresión Logística	Metodología Bayesiana
1, 2 y 3	67.0%	71.7%	70.3%	76.7%
4	33.0%	28.3%	29.7%	23.3%

Cuadro de porcentaje de clasificación original contra predicho (Análisis Discriminante)

Grupo Original	Predicho		
	1, 2 y 3	4	Total
1, 2 y 3	89.0%	11.0%	100.0%
4	36.7%	63.3%	100.0%

Cuadro de porcentaje de clasificación original contra predicho (Regresión logística)

Grupo Original	Predicho		
	1, 2 y 3	4	Total
1, 2 y 3	89.0%	11.0%	100.0%
4	32.2%	67.8%	100.0%

Cuadro de porcentaje de clasificación original contra predicho (Metodología Bayesiana)

Grupo Original	Predicho		
	1, 2 y 3	4	Total
1, 2 y 3	92.6%	7.4%	100.0%
4	44.4%	55.6%	100.0%

Una vez más comparando los análisis se aprecia que al realizarlos de manera separada, la primera agrupación proporciona los mejores porcentajes de clasificación, sin embargo en este caso, la clasificación mas equilibrada parece ser la que junta los primeros dos grupos por una parte y los siguientes dos grupos por otra parte.

Se tiene entonces que para la propuesta del Comité, las técnicas pueden generar agrupaciones que en conjunto lleven a replicar la clasificación original de una mejor manera que para la segunda alternativa, sin embargo, esto produce que las distribuciones porcentuales de hogares por grupo se distorsionen (lo cual no ocurre con la segunda alternativa) y que al interior de los grupos haya errores de clasificación mucho más radicales que los que ocurren en la segunda alternativa, para la cual los cambios se dan en las fronteras, principalmente al utilizar el análisis discriminante como técnica de clasificación. Es importante señalar sin embargo, que el tercer grupo se distorsiona mucho cuando se aplican las técnicas a la segunda alternativa. Adicionalmente, la segunda alternativa comete más errores cuando intenta agrupar los extremos que la primera alternativa.

A continuación se realizará la aplicación de las reglas de clasificación obtenidas a partir de ambas propuestas iniciales y para las tres técnicas analizadas.

e. Validación de las propuestas por medio de la información del año 2002

Las reglas de asignación obtenidas por discriminante, regresión logística y metodología bayesiana utilizando ambas alternativas de clasificación inicial a partir de los datos del año 2000, se aplican a los hogares captados en la ENIGH del año 2002 con el objetivo de verificar que la eficiencia en términos de la clasificación no disminuye en un lapso no muy amplio de tiempo. Un objetivo para realizar esto es determinar si, al encontrar una regla de clasificación para los datos de 2002, es válido utilizarla para agrupar individuos después de dos años, es decir, al momento en que este trabajo se está concluyendo. Es importante señalar que los datos referentes a ingresos se deflactan de acuerdo con el índice nacional de precios al consumidor.

Las reglas obtenidas a partir de la primera alternativa generan una clasificación para 2002 con las mismas problemáticas que la clasificación generada para 2000, tienen similares porcentajes de clasificación global, eliminan al grupo 2 y clasifican mejor en los extremos, y generan una distorsión en la distribución porcentual original de los hogares por grupos, como se puede apreciar en los siguientes cuadros.

Porcentaje de clasificación total correcta por método utilizado	
Método	Porcentaje
Análisis Discriminante	60.1%
Regresión logística	60.8%
Metodología Bayesiana	58.2%

Distribución de la clasificación por grupos según metodología				
Grupo	Original	Análisis Discriminante	Regresión Logística	Metodología Bayesiana
1	25.1%	23.8%	26.0%	22.5%
2	7.9%	0.0%	0.0%	0.0%
3	22.8%	17.8%	14.1%	6.5%
4	44.2%	58.4%	59.9%	71.1%

Cuadro de porcentaje de clasificación original contra predicho (Análisis Discriminante)					
Grupo Original	Predicho				
	1	2	3	4	Total
1	62.5%	0.0%	20.4%	17.1%	100.0%
2	33.0%	0.0%	28.9%	38.1%	100.0%
3	15.5%	0.0%	27.5%	57.0%	100.0%
4	4.5%	0.0%	9.3%	86.2%	100.0%

Cuadro de porcentaje de clasificación original contra predicho (Regresión Logística)					
Grupo Original	Predicho				
	1	2	3	4	Total
1	66.3%	0.0%	15.0%	18.8%	100.0%
2	36.7%	0.0%	24.7%	38.6%	100.0%
3	18.6%	0.0%	23.3%	58.1%	100.0%
4	5.1%	0.0%	6.9%	88.0%	100.0%

Cuadro de porcentaje de clasificación original contra predicho (Metodología Bayesiana)					
Grupo Original	Predicho				
	1	2	3	4	Total
1	59.9%	0.0%	9.4%	30.7%	100.0%
2	30.7%	0.0%	12.5%	56.8%	100.0%
3	14.4%	0.0%	8.5%	77.1%	100.0%
4	4.0%	0.0%	2.7%	93.3%	100.0%

En el caso de las reglas de asignación obtenidas a partir de la segunda alternativa de clasificación inicial también se observa que los problemas encontrados en la clasificación obtenida para 2000 se replican en la de 2002, respetando más las proporciones por grupo, clasificando mal a los grupos uno y tres, pero reclasificando de manera menos radical a los hogares mal clasificados con respecto a los resultados a partir de la primera alternativa, y destacándose nuevamente que la técnica que más efectiva resulta es el análisis discriminante como se puede apreciar en las siguientes tablas.

Porcentaje de clasificación total correcta por método utilizado	
Método	Porcentaje
Análisis Discriminante	55.9%
Regresión logística	56.7%
Metodología Bayesiana	52.6%

Distribución de la clasificación por grupos según metodología				
Grupo	Original	Análisis Discriminante	Regresión Logística	Metodología Bayesiana
1	11.6%	11.4%	6.0%	4.2%
2	33.5%	30.5%	38.9%	38.0%
3	26.6%	27.7%	21.2%	17.0%
4	28.3%	30.4%	33.9%	40.8%

Cuadro de porcentaje de clasificación original contra predicho (Análisis Discriminante)					
Grupo Original	Predicho				
	1	2	3	4	Total
1	57.9%	36.6%	4.7%	0.8%	100.0%
2	12.6%	53.3%	26.5%	7.6%	100.0%
3	1.5%	24.3%	43.7%	30.5%	100.0%
4	0.1%	6.9%	23.5%	69.5%	100.0%

Cuadro de porcentaje de clasificación original contra predicho (Regresión Logística)					
Grupo Original	Predicho				
	1	2	3	4	Total
1	35.1%	60.2%	3.5%	1.1%	100.0%
2	5.4%	65.2%	20.8%	8.5%	100.0%
3	0.2%	29.9%	34.7%	35.2%	100.0%
4	0.0%	7.6%	16.2%	76.3%	100.0%

Cuadro de porcentaje de clasificación original contra predicho (Metodología Bayesiana)					
Grupo Original	Predicho				
	1	2	3	4	Total
1	27.0%	66.7%	4.3%	2.0%	100.0%
2	3.0%	61.2%	19.7%	16.1%	100.0%
3	0.2%	26.9%	24.8%	48.1%	100.0%
4	0.0%	9.3%	11.6%	79.2%	100.0%

En cualquier caso ambas reglas resultan ser consistentes y robustas con respecto al tiempo. Debido a estos resultados hay elementos para decidir que la regla de asignación obtenida a partir de la segunda alternativa de clasificación inicial y aplicando la técnica de análisis discriminante es la más recomendable para asignar un nuevo hogar a un grupo dadas sus características.

Es importante reiterar que los análisis están partiendo de la premisa de que las variables indicadoras contienen suficiente información para generar por sí solas una definición del concepto de pobreza, lo cual no necesariamente es cierto.

f. Características comparativas de los hogares según el grupo al que pertenecen

Una vez que se ha corroborado la validez en la aplicación de las técnicas de agrupación en el año 2002, se presenta esta sección con el objetivo de observar las diferencias en las características de los hogares dependiendo del grupo al que pertenecen. Esto se realizará a partir de la clasificación generada por la regla de asignación obtenida utilizando la segunda alternativa de clasificación inicial y la técnica de análisis discriminante. Se destaca que en su momento el Comité Técnico para la Medición de la Pobreza realizó un análisis descriptivo similar y con ello justificó la validez de su agrupación.

Al analizar el tamaño de la localidad en que habitan los hogares se observa que a mayor tamaño existe una menor probabilidad de pertenecer a un grupo en condición de pobreza. Esto suena razonable dadas las oportunidades de acceso a servicios.

Características generales					
	Grupo				
	1	2	3	4	Total
Tamaño de la localidad de residencia					
Loc. de 100,000 hab. y más	1.9%	20.1%	57.2%	78.7%	49.6%
Loc. de 15,000 a 99,000 hab.	3.5%	15.1%	19.4%	11.9%	14.2%
Loc. de 2,500 a 14,999 hab.	12.8%	18.8%	13.9%	6.4%	12.6%
Loc de menos de 2,500 hab.	81.9%	45.9%	9.6%	3.0%	23.7%

Servicios					
	Grupo				
	1	2	3	4	Total
Drenaje	11.7%	50.3%	90.1%	98.8%	75.3%
Recolección pública de basura	20.2%	66.3%	93.4%	97.7%	81.1%
Luz	79.6%	98.5%	99.9%	100.0%	97.8%
Disponibilidad de agua					
sin agua	43.4%	19.9%	2.8%	0.3%	10.2%
agua en el terreno	53.0%	61.5%	18.3%	1.5%	27.6%
agua en la vivienda	3.5%	18.6%	78.9%	98.2%	62.2%

En el cuadro anterior se destaca que los grupos de los extremos se marcan muy notoriamente como grupos radicalmente opuestos, pues mientras en el primero prácticamente no se tiene acceso a drenaje o a recolección de basura, en el último grupo un hogar que no tenga estos servicios es la excepción. La disponibilidad de agua marca una línea muy

importante de distinción ya que en el caso del primer grupo casi la mitad de los hogares no tienen este indispensable insumo de vida.

El material de los techos en el hogar refleja, como se puede apreciar en la opción tabique, un claro ordenamiento en los cuatro grupos, al igual que el material de los muros. Estos dos indicadores generalmente van relacionados dado que son los muros los que sostienen los techos, y un mal material en muros no puede sustentar un buen material de techos. Se observa en el primer grupo el uso aún frecuente de materiales no perdurables, como láminas, madera y adobe, materiales que prácticamente están en desuso en el último grupo.

Características de la vivienda - Material de los techos					
	Grupo				
	1	2	3	4	Total
Material de los techos					
Cartón, hule, tela, llantas, etc.	0.2%	0.0%	0.0%	0.0%	0.0%
Lámina de cartón	18.0%	5.3%	2.2%	0.2%	3.7%
Palma, tejamanil o madera	9.1%	3.2%	3.4%	4.8%	4.3%
Lámina metálica, fibra de vidrio, plástico o mica	38.2%	22.7%	8.4%	1.8%	12.6%
Carrizo, bambú y terrado	1.3%	2.0%	1.2%	0.5%	1.1%
Lámina de asbesto	12.0%	14.7%	7.5%	1.7%	7.9%
Teja	11.9%	9.5%	1.6%	0.5%	4.3%
Panel de concreto	0.0%	0.2%	0.8%	1.9%	0.9%
Concreto monolítico	0.3%	0.4%	0.8%	1.1%	0.8%
Tabique, ladrillo, tabicón o loza de concreto	7.1%	37.8%	68.6%	81.8%	59.4%
Block	0.2%	0.2%	0.6%	0.6%	0.4%
Vigueta y poliuretano, viguieta y bovedilla, vigueta y cuña	1.2%	3.3%	4.6%	4.9%	4.1%

Características de la vivienda - Material de los muros					
	Grupo				
	1	2	3	4	Total
Material de los muros					
Cartón, hule, tela, llantas, etc.	0.0%	0.1%	0.0%	0.0%	0.0%
Lámina de cartón	0.9%	0.9%	0.2%	0.0%	0.4%
Carrizo, bambú, palma o tejamanil	5.1%	0.6%	0.1%	0.0%	0.6%
Embarro o bejaraque	13.7%	2.6%	0.0%	0.0%	1.9%
Lámina de asbesto	0.0%	0.0%	0.0%	0.0%	0.0%
Lámina metálica, fibra de vidrio, plástico o mica	2.6%	0.7%	0.1%	0.0%	0.5%
Madera	26.2%	7.8%	2.4%	1.6%	5.6%
Vidrio o cristal	0.0%	0.0%	0.0%	0.0%	0.0%
Panel de concreto	0.0%	0.2%	0.6%	1.1%	0.6%
Concreto monolítico	0.0%	0.0%	0.2%	1.1%	0.5%
Adobe	19.9%	15.4%	7.5%	2.3%	8.9%
Tabique, ladrillo, tabicón o block	26.7%	68.5%	87.4%	92.8%	78.9%
Piedra o cemento (incluye cantera)	4.2%	2.9%	1.4%	1.0%	1.9%
Otros materiales	0.6%	0.2%	0.1%	0.0%	0.2%

El material del piso muestra una clara distinción entre los grupos, teniendo en el primer grupo aún un muy alto porcentaje de viviendas con piso de tierra, con la insalubridad que esto conlleva.

Características de la vivienda - Material del piso					
	Grupo				
	1	2	3	4	Total
Material del piso					
Tierra	72.0%	11.9%	0.9%	0.2%	9.7%
Cemento o firme	25.5%	80.6%	75.9%	20.0%	53.2%
Madera, mosaico, loseta de concreto, loseta de plástico	2.6%	7.5%	23.2%	79.9%	36.9%

Los enseres son un reflejo claro de la capacidad adquisitiva de los hogares dentro de cada grupo, no es fácil para un hogar del primer grupo tener acceso a un refrigerador o una lavadora, ni siquiera a una licuadora, y por supuesto que ni contemplar el tener una computadora, bien que está restringido al último grupo.

Posesión de enseres					
	Grupo				
	1	2	3	4	Total
Enser					
Radio	19.3%	23.4%	29.2%	36.3%	29.2%
Televisión	45.5%	84.8%	97.1%	98.3%	89.7%
Computadora	0.1%	0.5%	0.0%	38.0%	13.7%
Refrigerador	13.1%	57.7%	88.9%	97.3%	76.6%
Licuadora	31.5%	74.1%	91.7%	96.6%	83.4%
Lavadora	5.7%	27.9%	66.6%	86.0%	57.4%
Vehículo	0.6%	2.9%	8.4%	62.3%	25.3%
Teléfono	0.6%	11.0%	42.7%	72.7%	40.8%

Ya en algún momento del presente trabajo se había planteado la cuestión de las diversas problemáticas en el uso de leña como combustible, sin embargo no había sido muy clara la importancia de este rubro, en el siguiente cuadro se puede apreciar que es justamente el primer grupo el que utiliza prácticamente en su totalidad la leña como combustible, cuando la gran mayoría de las personas utiliza actualmente gas.

Combustible que utiliza					
	Grupo				
	1	2	3	4	Total
Tipo de combustible					
Leña	94.7%	24.4%	0.3%	0.3%	15.0%
Carbón	0.3%	0.5%	0.0%	0.0%	0.2%
Petróleo	0.7%	0.1%	0.0%	0.0%	0.1%
Electricidad	0.6%	0.5%	0.1%	0.8%	0.5%
Gas	2.8%	73.2%	99.2%	98.1%	83.3%
Otros	0.4%	0.2%	0.0%	0.3%	0.2%
No utiliza combustible	0.6%	1.1%	0.4%	0.5%	0.6%

Es claro que el primer grupo es el más aquejado por la pobreza, pero a ésto hay que agregarle problemáticas que vienen de cuestiones costumbristas, y que agravan en mayor medida su situación, como la baja planificación familiar, el arraigo a sus lugares de origen, lo cual produce altos niveles de hacinamiento, baja escolaridad (por el bajo acceso a servicios educativos), altos niveles de dependencia y por ende menores ingresos per cápita que generan un círculo vicioso del que es difícil escapar.

Medias de algunos indicadores directos e indirectos					
	Grupo				
	1	2	3	4	Total
Variable					
Número de cuartos para dormir	1.6	2.2	3.0	3.9	3.0
Tamaño del hogar	5.4	4.4	3.9	3.8	4.1
Índice de hacinamiento (personas por cuarto)	3.9	2.4	1.5	1.1	1.8
Años de escolaridad del jefe del hogar	2.3	4.0	5.5	11.3	6.9
Índice de dependencia demográfica (personas en los grupos de edad 0 a 15 años y 65 años en adelante, entre personas en el grupo de 16 a 64 años)	1.5	1.1	0.7	0.5	0.8
Ingreso mensual del hogar	1995.6	3472.6	5392.2	12748.9	7174.1
Ingreso mensual per cápita del hogar	386.9	850.2	1518.3	3937.0	2092.4
Gasto mensual per cápita	344.4	749.1	1273.9	3172.4	1721.2

Es importante realizar una sensibilización de los problemas que viven los hogares de los primeros grupos, tomando en cuenta que cada individuo del primer grupo tiene que vivir con alrededor de 11 pesos diarios en un momento en el que un kilo de tortillas cuesta en promedio entre 5.50 y 6 pesos, un litro de leche (subsidiada) 3.50 pesos y una coca cola de trescientos cincuenta y cinco mililitros retornable 3 pesos.

Finalmente es importante destacar a manera comparativa, que en esta propuesta de clasificación se está encontrando un grupo de hogares que está incluido dentro de los hogares en pobreza alimentaria, es decir que no tienen para comer, pero cuya problemática va más allá (sí, aún puede ir mas allá del no tener para comer), pues no solo es un grupo con hambre, es un grupo con condiciones de vida infrahumanas, que es necesario atender de manera prioritaria.

Conclusiones

En lo referente a los objetivos planteados originalmente se concluye que la propuesta de clasificación de la población por su condición de pobreza elaborada por el Comité Técnico para la Medición de la Pobreza puede tener problemas al momento de ser replicada por medio de información multidimensional, lo cual indica que está restringida en cuanto a su definición, apegándose de manera estricta al concepto de pobreza por falta de ingresos.

No existe una clasificación que notoriamente descarte a la del Comité, sin embargo los datos multivariados por sí solos plantean la posibilidad de generar una clasificación que proviene de una definición de pobreza que los mismos datos sugieren. Esta clasificación alternativa puede replicarse de mejor manera con técnicas estadísticas, y refleja una forma diferente de agrupar a los hogares. Al validarla muestra un muy claro ordenamiento, lo cual implica la posibilidad de considerarla.

En lo que respecta a la propuesta de una regla de agrupación es importante señalar que durante la elaboración del proyecto se encontraron elementos discordantes para definir una técnica como óptima, puesto que para algunas agrupaciones una técnica era la mejor, y la misma técnica podía ser la menos útil para otras agrupaciones. Sin embargo en la mayoría de los casos se observó una consistencia entre los resultados de las técnicas cuando se aplicaron a la misma agrupación inicial.

Es claro que existe una compleja problemática cuando se trata de resolver un problema de agrupación como el que se planteó originalmente, puesto que distintos resultados no necesariamente están equivocados, dado que en muchos elementos se incorporan grados de subjetividad. Sin embargo, en el presente trabajo se observaron consistencias que conllevan a pensar en la factibilidad de involucrar muchos indicadores (en lugar de solo uno) para determinar la clasificación de la población por su condición de pobreza. Se seguirán encontrando diferencias, sin embargo mientras los indicadores que se utilicen para que por sí solos definan a la pobreza sean adecuados para dicho fin y sean suficientes, se estará en condiciones de aseverar que la clasificación es mejor cada vez.

Diversos autores han reiterado la necesidad de diseñar instrumentos de captación de información adecuados para medir las condiciones de vida de los individuos, en la medida en que estos instrumentos se lleven a la práctica, y que los indicadores que capten sean más completos en términos de reflejar la definición de la pobreza en la mayoría de sus

dimensiones, será de mayor utilidad y viabilidad aplicar las metodologías expuestas en el presente trabajo para obtener clasificaciones mucho más claras, y reglas de asignación de mayor robustez y eficiencia, reduciendo los errores de clasificación y permitiéndo que las estrategias de política social lleguen a quien deben llegar.

Finalmente, a manera de colofón se expresa una opinión personal, que consiste en que todo individuo, por el simple hecho de haber nacido, tiene derecho a llevar una vida digna, dentro de su entorno, con respeto a los demás, pero con igualdad, y debe ser obligación de todos garantizar esos derechos, y responsabilidad de cada individuo el uso de dichos derechos para bien.

Bibliografía

[1] Aarts, E. y Korst, J. (1989). *Simulated Annealing and Boltzmann Machines.* John Wiley Series.

[2] Agresti, A. (1990). *Categorical Data Análisis.* John Wiley & Sons. University of Florida.

[3] Atkinson, A.B. (1970). On Measurement of Inequality. *Journal of Economic Theory,* 2.

[4] Bermudez, J. D., Bernardo, J. M. y Sendra, M. (1987). Classification Problems in Education. *The Statistician.* 36, pp. 107-113.

[5] Bernardo, J. M. (1992). Simulated Annealing in Bayesian Theory. *Computational Statistics.* 1 Springer-Verlag.

[6] Bernardo, J. M. (1988). Bayesian Linear Probabilistic Classification. *Statistical decision Theory and Related Topics IV.* Springer-Verlag. 1 pp. 151-62.

[7] Bersimas, D. y Tsitsiklis, J. (1993). Simulated Annealing *Statistical Science.* 8-1 pp. 10-15.

[8] Bishop, Y. M., Fienberg, S. E. y Holland, P. W. (1975). *Discrete Multivariate Analysis.* MIT Press.

[9] Comité Técnico Para la Medición de la Pobreza. (2002). *La Medición de la Pobreza en México al Año 2000.* Subsecretaría de Prospectiva. Planeación y Evaluación de la Secretaría de Desarrollo Social.

[10] Comité Técnico Para la Medición de la Pobreza. (2002). *Medición de la Pobreza. variantes metodológicas y estimación preliminar.* serie: documentos de investigación. Secretaría de Desarrollo Social.

[11] Dawid, A. P. (1976). Properties of Diagnostic Data Distributions. *Biometrics* 32. pp. 647-658.

[12] Fienberg, S. E. (1979). *The Analysis of Cross-Classified Categorical Data.* The Massachusets Institute of Technology. MIT Press.

[13] Gallardo, R. y Osorio, J. (coords.). (1998). *Los Rostros de la Pobreza - El Debate. Tomo 1.* ITESO-IBERO.

[14] Gilbert, J y Gilbert, L. (1995). *Linear Algebra and Matrix Theory.* Academic Press.

[15] Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis.* Academic Press INC.

[16] Hair, J. F., Anderson, R. E., Tatham, R. L. y Black, W. C. (1995). *Multivariate Data Analysis.* Prentice Hall. New Jersey.

[17] Huberty, C. J. (1994). *Applied Discriminant Analysis.* John Wiley & Sons. New York.

[18] Jobson, J. D. (1992). *Applied Multivariate Data Analysis. Volume I: Regression and Experimental Design.* Springer-Verlag.

[19] Jobson, J. D. (1992). *Applied Multivariate Data Analysis. Volume II: Categorical and Multivariate Methods.* Springer-Verlag.

[20] Johnson, R. A. y Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis.* Prentice Hall.

[21] Kirkpatrick, S. Gelatt Jr., C. D. y Vecchi, M. P. (1983) Optimization by Simulated Annealing. *Science.* 220. pp. 671-680.

[22] Kleinbaum, D. (1992) *Logistic Regression. A Self-Learning Text* Springer Verlag.

[23] Krishnaiah, P. R. y Kanal, L. N. (1990) *Handbook of statistics 2: Classification Pattern Recognition and Reduction of Dimensionality* North-Holland.

[24] Knuth, D. E. (1984) *The texbook.* Addisson Wesley Publishing Company.

[25] Lebart, L., Morineau, A. y Warwick K. M. (1977). *Multivariate Descriptive Statistical Analysis. Correspondence Analysis and Related Techniques for Large Matrices.* John Wiley & Sons.

[26] Lee, P. M. (1989). *Bayesian Statistics: An Introduction.* Oxford University Press.

[27] López-Calva, L. F. y Rodtriguez Chamussy, L. (2004). *Muchos rostros. un solo espejo: restricciones para la medición multidimensional de la pobreza en México.* Working Paper. Universidad de las Américas.

[28] Press, W. H., Flannery, B., Teukolsky, S. A. y Vetrerling, W. T. (1988). *Numerical Recipes in C.* Cambridge University Press.

[29] Press, S. J. y Wilson, S. (1978). *Choosing Between Logistic Regression and Discriminant Analysis.* Journal of the American Statistical Association. Vol 73 No. 364.

[30] Ravallion, M. (1996). *Issues in Measuring and in Modeling Poverty.* Policy Research Working Paper. The World Bank.

[31] Robert, C. P. (2001). *The Bayesian Choice.* Springer-Verlag.

[32] Sen, A. K. (1997). *On Economic Inequality (expanded edition).* Clarendon Press. Oxford.

[33] Soto, H. (1999). *Algunas Técnicas para el Análisis de Datos Categóricos.* Tesis de licenciatura. ITAM.

[34] Titterington, D. M., Murray G. D., Murray, J. S., Spiegelhalter, D. J., Skene, A. M., Habbema, J. D. F. y Gelpke G. J. (1981). Comparison of Discrimination Techniques Applied to a Complex Data Set of Head Injured Patiens. *Journal of Royal Statistics Society.* 144. Part 2. pp. 145-175.

Apéndice I:

Demostración de la propuesta bayesiana de distribución predictiva

La distribución t de Student k-variada con α grados de libertad, parámetro de localización ϕ y matriz de precisión Ψ está dada por:

$$St_k(\underline{x}|\alpha,\phi,\Psi) = c[1+\frac{1}{\alpha}(\underline{x}-\phi)'\Psi(\underline{x}-\phi)]^{-\frac{\alpha+k}{2}}$$

donde:

$$c = \frac{\Gamma(\frac{1}{2}(\alpha+k))}{\Gamma(\frac{1}{2}\alpha)(\alpha\pi)^{\frac{k}{2}}}|\Psi|^{\frac{1}{2}} \quad ; \quad \alpha \in \Re^{+} \quad ; \quad x,\phi \in \Re^{k} \ ;$$

$$\Psi \in \Re^{kxk} \text{ simétrica y positiva definida ;}$$

$$E[\underline{x}] = \phi \text{ si } \alpha > 1 \ ;$$

$$V[\underline{x}] = \frac{\alpha}{\alpha-2}\Psi^{-1} \text{ si } \alpha > 2$$

La distribución Wishart k-variada con φ grados de libertad y matriz de precisión Ω está dada por:

$$W_k(V|\varphi,\Omega) = c|V|^{\frac{\varphi-(k+1)}{2}}exp\{\frac{-tr(V\Omega)}{2}\}$$

donde:

$$c = \frac{|\Omega|^{\frac{\varphi}{2}}}{2^{\frac{\varphi k}{2}}\pi^{\frac{k(k-1)}{4}}\prod_{i=1}^{k}\Gamma(\frac{\varphi-i+1}{2})} \ ;$$

$$V, \Omega \in \Re^{kxk} \text{ simétricas y positivas definidas ;}$$

$$\varphi \in \Re^{+} ;$$

$$E[V] = \varphi \Omega^{-1}$$

Algunas demostraciones y propiedades importantes:

1)

$$\sum_{j=1}^{n_i} (\underline{t}_{ij} - \underline{\mu}_i)' H_i (\underline{t}_{ij} - \underline{\mu}_i) =$$

$$\sum_{j=1}^{n_i} (\underline{t}_{ij} - \bar{\underline{t}}_i + \bar{\underline{t}}_i - \underline{\mu}_i)' H_i (\underline{t}_{ij} - \bar{\underline{t}}_i + \bar{\underline{t}}_i - \underline{\mu}_i) =$$

$$\sum_{j=1}^{n_i} (\underline{t}_{ij} - \bar{\underline{t}}_i)' H_i (\underline{t}_{ij} - \bar{\underline{t}}_i) + \sum_{j=1}^{n_i} (\underline{t}_{ij} - \bar{\underline{t}}_i)' H_i (\bar{\underline{t}}_i - \underline{\mu}_i) + \sum_{j=1}^{n_i} (\bar{\underline{t}}_i - \underline{\mu}_i)' H_i (\underline{t}_{ij} - \bar{\underline{t}}_i) + \sum_{j=1}^{n_i} (\bar{\underline{t}}_i - \underline{\mu}_i)' H_i (\bar{\underline{t}}_i - \underline{\mu}_i) =$$

$$\sum_{j=1}^{n_i} (\underline{t}_{ij} - \bar{\underline{t}}_i)' H_i (\underline{t}_{ij} - \bar{\underline{t}}_i) + n_i (\bar{\underline{t}}_i - \underline{\mu}_i)' H_i (\bar{\underline{t}}_i - \underline{\mu}_i) + (\sum_{j=1}^{n_i} \underline{t}_{ij} - n_i \bar{\underline{t}}_i)' H_i (\bar{\underline{t}}_i - \underline{\mu}_i) + (\bar{\underline{t}}_i - \underline{\mu}_i)' H_i (\sum_{j=1}^{n_i} \underline{t}_{ij} - n_i \bar{\underline{t}}_i) =$$

$$\sum_{j=1}^{n_i} (\underline{t}_{ij} - \bar{\underline{t}}_i)' H_i (\underline{t}_{ij} - \bar{\underline{t}}_i) + \sum_{j=1}^{n_i} (\bar{\underline{t}}_i - \mu_i)' H_i (\bar{\underline{t}}_i - \underline{\mu}_i) + (n_i \underline{\bar{t}} - n_i \bar{\underline{t}}_i)' H_i (\bar{\underline{t}}_i - \underline{\mu}_i) + (\bar{\underline{t}}_i - \underline{\mu}_i)' H_i (n_i \underline{\bar{t}} - n_i \bar{\underline{t}}_i) =$$

$$\sum_{j=1}^{n_i} (\underline{t}_{ij} - \bar{\underline{t}}_i)' H_i (\underline{t}_{ij} - \bar{\underline{t}}_i) + \sum_{j=1}^{n_i} (\bar{\underline{t}}_i - \underline{\mu}_i)' H_i (\bar{\underline{t}}_i - \underline{\mu}_i)$$

2)

$$tr(x) = x \text{ si } x \in \Re$$

3)

$$tr(AB) = tr(BA)$$

4)

$$\sum_{j=1}^{n_i} H_i(\underline{t}_{ij} - \bar{\underline{t}}_i)'(\underline{t}_{ij} - \bar{\underline{t}}_i) = H_i[\sum_{j=1}^{n_i} (\underline{t}_{ij} - \bar{\underline{t}}_i)'(\underline{t}_{ij} - \bar{\underline{t}}_i)]$$

$$= H_i[n_i S_i]$$

5)

$$(m+n)(\underline{\mu} - \frac{m\underline{x} + n\underline{y}}{m+n})'H(\underline{\mu} - \frac{m\underline{x} + n\underline{y}}{m+n}) + \frac{mn}{m+n}(\underline{x} - \underline{y})'H(\underline{x} - \underline{y}) =$$

$$(m+n)\underline{\mu}'H\underline{\mu} - \underline{\mu}'H(m\underline{x} + n\underline{y}) - (m\underline{x} + n\underline{y})'H\underline{\mu} + \frac{(m\underline{x} + n\underline{y})'H(m\underline{x} + n\underline{y})}{m+n} + \frac{mn}{m+n}\underline{x}'H\underline{x} - \frac{mn}{m+n}\underline{x}'H\underline{y} - \frac{mn}{m+n}\underline{y}'H\underline{x} + \frac{mn}{m+n}\underline{y}'H\underline{y} =$$

$$m\underline{\mu}'H\underline{\mu} + n\underline{\mu}'H\underline{\mu} - m\underline{\mu}'H\underline{x} - n\underline{\mu}'H\underline{y} - m\underline{x}'H\underline{\mu} - n\underline{y}'H\underline{\mu} + \frac{m^2}{m+n}\underline{x}'H\underline{x} + \frac{mn}{m+n}\underline{x}'H\underline{y}$$

$$+\frac{mn}{m+n}\underline{y}'H\underline{x} + \frac{n^2}{m+n}\underline{y}'H\underline{y} + \frac{mn}{m+n}\underline{x}'H\underline{x} - \frac{mn}{m+n}\underline{x}'H\underline{y} - \frac{mn}{m+n}\underline{y}'H\underline{x} + \frac{mn}{m+n}\underline{y}'H\underline{y} =$$

$$m\underline{\mu}'H\underline{\mu} + n\underline{\mu}'H\underline{\mu} - m\underline{\mu}'H\underline{x} - n\underline{\mu}'H\underline{y} - m\underline{x}'H\underline{\mu} - n\underline{y}'H\underline{\mu} + (\frac{m^2}{m+n} + \frac{mn}{m+n})\underline{x}'H\underline{x} + (\frac{n^2}{m+n} + \frac{mn}{m+n})\underline{y}'H\underline{y} =$$

$$m\underline{\mu}'H\underline{\mu} + n\underline{\mu}'H\underline{\mu} - m\underline{\mu}'H\underline{x} - n\underline{\mu}'H\underline{y} - m\underline{x}'H\underline{\mu} - n\underline{y}'H\underline{\mu} + (\frac{m(m+n)}{m+n})\underline{x}'H\underline{x} + (\frac{n(m+n)}{m+n})\underline{y}'H\underline{y} =$$

$$m\underline{\mu}'H\underline{\mu} + n\underline{\mu}'H\underline{\mu} - m\underline{\mu}'H\underline{x} - n\underline{\mu}'H\underline{y} - m\underline{x}'H\underline{\mu} - n\underline{y}'H\underline{\mu} + m\underline{x}'H\underline{x} + n\underline{y}'H\underline{y} =$$

$$(m\underline{x}'H\underline{x} - m\underline{x}'H\underline{\mu} - m\underline{\mu}'H\underline{x} + m\underline{\mu}'H\underline{\mu}) + (n\underline{y}'H\underline{y} - n\underline{y}'H\underline{\mu} - n\underline{\mu}'H\underline{y} + n\underline{\mu}'H\underline{\mu}) =$$

$$m(\underline{x} - \underline{\mu})'H(\underline{x} - \underline{\mu}) + n(\underline{y} - \underline{\mu})'H(\underline{y} - \underline{\mu})$$

6)

$$\begin{aligned} |V_l + (\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})'| &= |I_l + V^{-1}(\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})'| \\ &= 1 + (\underline{x} - \underline{\mu})'V^{-1}(\underline{x} - \underline{\mu}) \end{aligned}$$

LA DEMOSTRACIÓN

$$\int_{H_i}\int_{\underline{\mu}_i} |H_i|^{-\frac{\nu}{2}}[\prod_{j=1}^{n_i}(2\pi)^{-\frac{l}{2}}|H_i|^{\frac{1}{2}}exp[-\frac{1}{2}(\underline{t}_{ij}-\underline{\mu}_i)'H_i(\underline{t}_{ij}-\underline{\mu}_i)](2\pi)^{-\frac{l}{2}}|H_i|^{\frac{1}{2}}exp[-\frac{1}{2}(\underline{t}-\underline{\mu}_i)'H_i(\underline{t}-\underline{\mu}_i)]d\underline{\mu}_i dH_i$$

$$= \int_{H_i}\int_{\underline{\mu}_i} |H_i|^{\frac{(n_i-\nu+1)}{2}}[(2\pi)^{-\frac{l(n_i+1)}{2}}]exp[-\frac{1}{2}[\sum_{j=1}^{n_i}(\underline{t}_{ij}-\underline{\mu}_i)'H_i(\underline{t}_{ij}-\underline{\mu}_i)+(\underline{t}-\underline{\mu}_i)'H_i(\underline{t}-\underline{\mu}_i)]]d\underline{\mu}_i dH_i$$

por 1)

$$= \int_{H_i}\int_{\underline{\mu}_i} |H_i|^{\frac{(n_i-\nu+1)}{2}}[(2\pi)^{-\frac{l(n_i+1)}{2}}]exp[-\frac{1}{2}\sum_{j=1}^{n_i}((\underline{t}_{ij}-\bar{\underline{t}}_i)'H_i(\underline{t}_{ij}-\bar{\underline{t}}_i))-\frac{1}{2}\sum_{j=1}^{n_i}((\bar{\underline{t}}_i-\underline{\mu}_i)'H_i(\bar{\underline{t}}_i-\underline{\mu}_i))-\frac{1}{2}(\underline{t}-\underline{\mu}_i)'H_i(\underline{t}-\underline{\mu}_i)]d\underline{\mu}_i dH_i$$

por 2) y 3)

$$= \int_{H_i}\int_{\underline{\mu}_i} |H_i|^{\frac{(n_i-\nu+1)}{2}}[(2\pi)^{-\frac{l(n_i+1)}{2}}]exp[-\frac{1}{2}\sum_{j=1}^{n_i}(H_i(\underline{t}_{ij}-\bar{\underline{t}}_i)(\underline{t}_{ij}-\bar{\underline{t}}_i)')-\frac{1}{2}\sum_{j=1}^{n_i}((\bar{\underline{t}}_i-\underline{\mu}_i)'H_i(\bar{\underline{t}}_i-\underline{\mu}_i))-\frac{1}{2}(\underline{t}-\underline{\mu}_i)'H_i(\underline{t}-\underline{\mu}_i)]d\underline{\mu}_i dH_i$$

por 4)

$$= \int_{H_i}\int_{\underline{\mu}_i} |H_i|^{\frac{(n_i-\nu+1)}{2}}[(2\pi)^{-\frac{l(n_i+1)}{2}}]exp[-\frac{1}{2}H_i[n_iS_i]]exp[-\frac{1}{2}\sum_{j=1}^{n_i}(\bar{\underline{t}}_i-\underline{\mu}_i)'H_i(\bar{\underline{t}}_i-\underline{\mu}_i)]exp[-\frac{1}{2}(\underline{t}-\underline{\mu}_i)'H_i(\underline{t}-\underline{\mu}_i)]d\underline{\mu}_i dH_i$$

por 2) y 3)

$$= \int_{H_i}\int_{\underline{\mu}_i} |H_i|^{\frac{(n_i-\nu+1)}{2}}[(2\pi)^{-\frac{l(n_i+1)}{2}}]exp[-\frac{n_i}{2}tr(H_iS_i)]exp[-\frac{1}{2}\sum_{j=1}^{n_i}(\underline{\bar{t}}_i-\underline{\mu}_i)'H_i(\underline{\bar{t}}_i-\underline{\mu}_i)]exp[-\frac{1}{2}(\underline{t}-\underline{\mu}_i)'H_i(\underline{t}-\underline{\mu}_i)]d\underline{\mu}_i dH_i$$

$$= \int_{H_i}\int_{\underline{\mu}_i} |H_i|^{\frac{(n_i-\nu+1)}{2}}[(2\pi)^{-\frac{l(n_i+1)}{2}}]exp[-\frac{n_i}{2}tr(H_iS_i)]exp[-\frac{n_i}{2}(\underline{\bar{t}}_i-\underline{\mu}_i)'H_i(\underline{\bar{t}}_i-\underline{\mu}_i)-\frac{1}{2}(\underline{t}-\underline{\mu}_i)'H_i(\underline{t}-\underline{\mu}_i)]d\underline{\mu}_i dH_i$$

por 5)

$$= \int_{H_i}\int_{\underline{\mu}_i} |H_i|^{\frac{(n_i-\nu+1)}{2}}[(2\pi)^{-\frac{l(n_i+1)}{2}}]exp[-\frac{n_i}{2}tr(H_iS_i)]exp[-\frac{n_i+1}{2}(\underline{\mu}_i-(\frac{n_i\underline{\bar{t}}_i+\underline{t}}{n_i+1}))'H_i(\underline{\mu}_i-(\frac{n_i\underline{\bar{t}}_i+\underline{t}}{n_i+1}))]exp[-\frac{n_i}{2(n_i+1)}(\underline{\bar{t}}_i-\underline{t})'H_i(\underline{\bar{t}}_i-\underline{t})]d\underline{\mu}_i dH_i$$

por 2) y 3)

$$= \int_{H_i}\int_{\underline{\mu}_i} |H_i|^{\frac{(n_i-\nu+1)}{2}}[(2\pi)^{-\frac{l(n_i+1)}{2}}]exp[-\frac{n_i}{2}tr(H_iS_i)]exp[-\frac{n_i+1}{2}(\underline{\mu}_i-(\frac{n_i\underline{\bar{t}}_i+\underline{t}}{n_i+1}))'H_i(\underline{\mu}_i-(\frac{n_i\underline{\bar{t}}_i+\underline{t}}{n_i+1}))]exp[-\frac{n_i}{2(n_i+1)}tr(H_i(\underline{\bar{t}}_i-\underline{t})(\underline{\bar{t}}_i-\underline{t})')]d\underline{\mu}_i dH_i$$

$$= \int_{H_i}\int_{\underline{\mu}_i} |H_i|^{\frac{(n_i-\nu+1)}{2}}[(2\pi)^{-\frac{l(n_i+1)}{2}}]exp[-\frac{1}{2}H_i(n_iS_i+\frac{n_i}{(n_i+1)}(\underline{\bar{t}}_i-\underline{t})(\underline{\bar{t}}_i-\underline{t})')]exp[-\frac{n_i+1}{2}(\underline{\mu}_i-(\frac{n_i\underline{\bar{t}}_i+\underline{t}}{n_i+1}))'H_i(\underline{\mu}_i-(\frac{n_i\underline{\bar{t}}_i+\underline{t}}{n_i+1}))]d\underline{\mu}_i dH_i$$

reacomodando términos se tiene:

$$= \int_{H_i} |H_i|^{\frac{(n_i-\nu)}{2}}[(2\pi)^{-\frac{ln_i}{2}}]exp[-\frac{1}{2}H_i(n_iS_i+\frac{n_i}{(n_i+1)}(\underline{\bar{t}}_i-\underline{t})(\underline{\bar{t}}_i-\underline{t})')]\int_{\underline{\mu}_i}|H_i|^{\frac{(1)}{2}}[(2\pi)^{-\frac{l}{2}}]exp[-\frac{n_i+1}{2}(\underline{\mu}_i-(\frac{n_i\underline{\bar{t}}_i+\underline{t}}{n_i+1}))'H_i(\underline{\mu}_i-(\frac{n_i\underline{\bar{t}}_i+\underline{t}}{n_i+1}))]d\underline{\mu}_i dH_i$$

de donde en la integral sobre $\underline{\mu}_i$ se identifica a una distribución Normal l-variada, para la cual se sabe que dicha integral es 1. por tanto solamente resta:

$$= \int_{H_i} |H_i|^{\frac{(n_i-\nu)}{2}} [(2\pi)^{-\frac{ln_i}{2}}] exp[-\frac{1}{2} H_i (n_i S_i + \frac{n_i}{(n_i+1)} (\bar{\underline{t}}_i - \underline{t})(\bar{\underline{t}}_i - \underline{t})')] dH_i$$

el integrando tiene una forma aproximada a la de una distribución Wishart l-variada, por lo que se completará dicha distribución. Para ello se definirán los parámetros de la siguiente manera.

Para una distribución Wishart $W_k(V|\varphi, \Omega)$, sean:

$$V = H_i \quad , \quad \varphi = n_i - \nu + l + 1 \quad , \quad \Omega = n_i S_i + \frac{n_i}{(n_i+1)} (\bar{\underline{t}}_i - \underline{t})(\bar{\underline{t}}_i - \underline{t})'$$

y una constante de proporcionalidad:

$$k = \frac{|\Omega|^{\frac{\varphi}{2}}}{d}$$

con:

$$d = \pi^{\frac{l(l+1)}{4}} \prod_{j=1}^{l} \Gamma(\frac{1}{2}(\varphi - l + 1))$$

Por tanto la integral se puede reescribir como:

$$[(2\pi)^{-\frac{ln_i}{2}}]\int_{H_i}|H_i|^{\frac{(\varphi-(l+1))}{2}}exp[-\frac{1}{2}H_i\Omega]dH_i$$

por 3)

$$=[(2\pi)^{-\frac{ln_i}{2}}]\int_{H_i}|H_i|^{\frac{(\varphi-(l+1))}{2}}exp[-\frac{tr(H_i\Omega)}{2}]dH_i$$

Se identifica a la distribución Wishart en la integral, se completan sus términos:

$$=[(2\pi)^{-\frac{ln_i}{2}}]k^{-1}\int_{H_i}k|H_i|^{\frac{(\varphi-(l+1))}{2}}exp[-\frac{tr(H_i\Omega)}{2}]dH_i$$

De donde se tiene:

$$=[(2\pi)^{-\frac{ln_i}{2}}]k^{-1}$$

$$=[(2\pi)^{-\frac{ln_i}{2}}]d(|\Omega|^{-\frac{\varphi}{2}})$$

El término d es constante al igual que el término $[(2\pi)^{-\frac{ln_i}{2}}]$, por tanto se tiene:

$$[(2\pi)^{-\frac{ln_i}{2}}]d(|\Omega|^{-\frac{\varphi}{2}})\propto|\Omega|^{-\frac{\varphi}{2}}$$

Sustituyendo el valor de Ω se tiene:

$$|n_iS_i+\frac{n_i}{(n_i+1)}(\underline{\bar{t}}_i-\underline{t})(\underline{\bar{t}}_i-\underline{t})'|^{-\frac{\varphi}{2}}$$

Sustituyendo el valor de φ y factorizando términos se tiene:

$$|n_i(S_i + \frac{1}{(n_i+1)}(\bar{\underline{t}}_i - \underline{t})(\bar{\underline{t}}_i - \underline{t})')|^{-\frac{n_i-\nu+l+1}{2}}$$

de nuevo por ser n_i constante se tiene que:

$$|n_i(S_i + \frac{1}{(n_i+1)}(\bar{\underline{t}}_i - \underline{t})(\bar{\underline{t}}_i - \underline{t})')|^{-\frac{n_i-\nu+l+1}{2}} \propto |S_i + \frac{1}{(n_i+1)}(\bar{\underline{t}}_i - \underline{t})(\bar{\underline{t}}_i - \underline{t})'|^{-\frac{n_i-\nu+l+1}{2}}$$

por 6)

$$\propto [1 + \frac{1}{(n_i+1)}(\bar{\underline{t}}_i - \underline{t})' S_i^{-1}(\bar{\underline{t}}_i - \underline{t})]^{-\frac{n_i-\nu+l+1}{2}}$$

y reacomodando términos:

$$\propto [1 + \frac{1}{(n_i-\nu+1)}(\bar{\underline{t}}_i - \underline{t})' \frac{n_i-\nu+1}{(n_i+1)} S_i^{-1}(\bar{\underline{t}}_i - \underline{t})]^{-\frac{n_i-\nu+l+1}{2}}$$

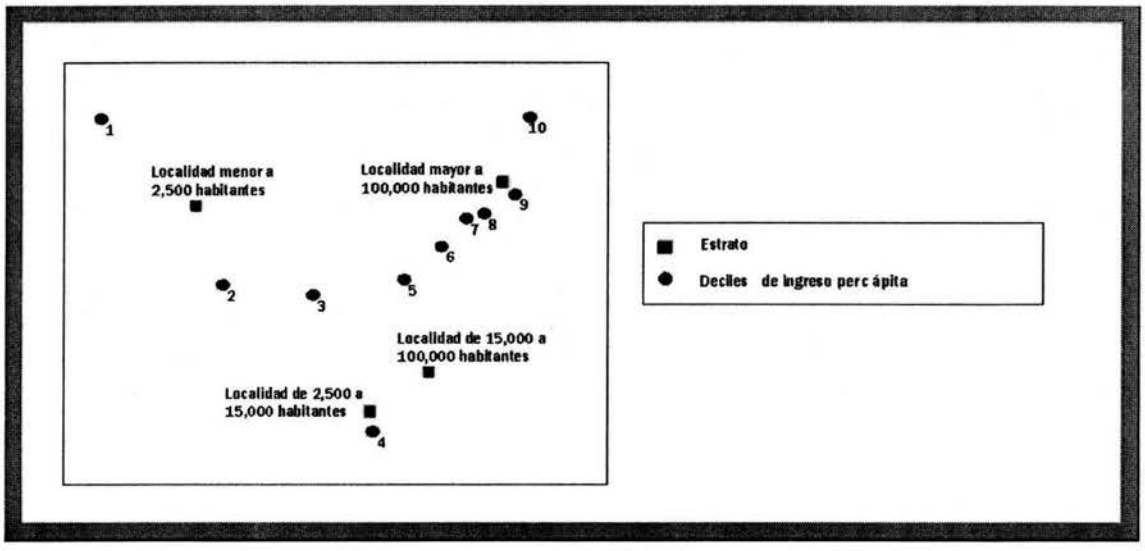
De donde se identifica el kernel de una distribución t de Student l-variada ($St_k(\underline{x}|\alpha, \phi, \Psi)$) con los siguientes parámetros:

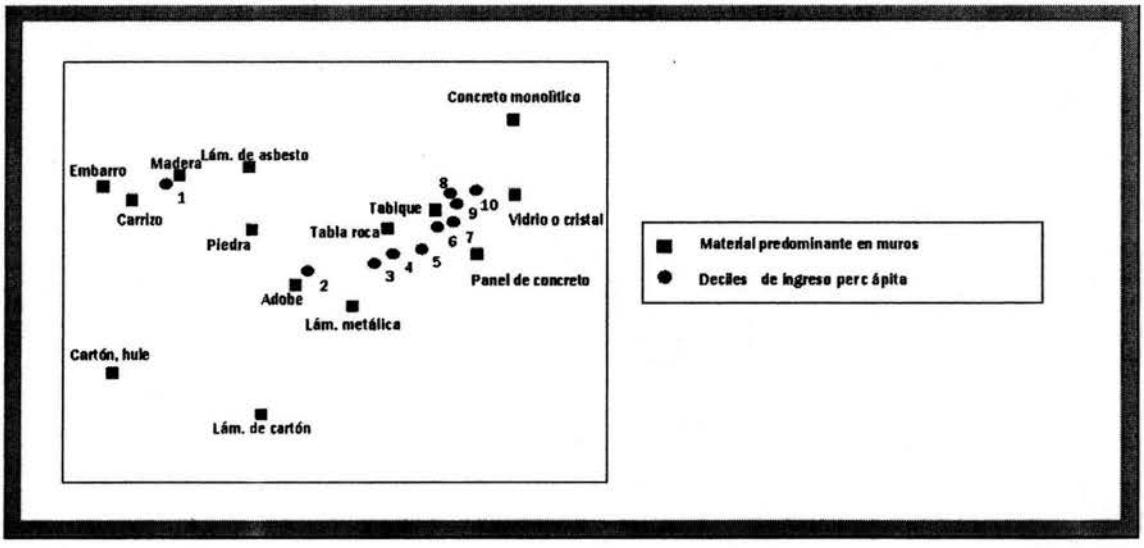
$$\alpha = n_i - \nu + 1 \quad , \quad \phi = \bar{\underline{t}}_i \quad , \quad \Psi = \frac{n_i-\nu+1}{(n_i+1)} S_i^{-1}$$

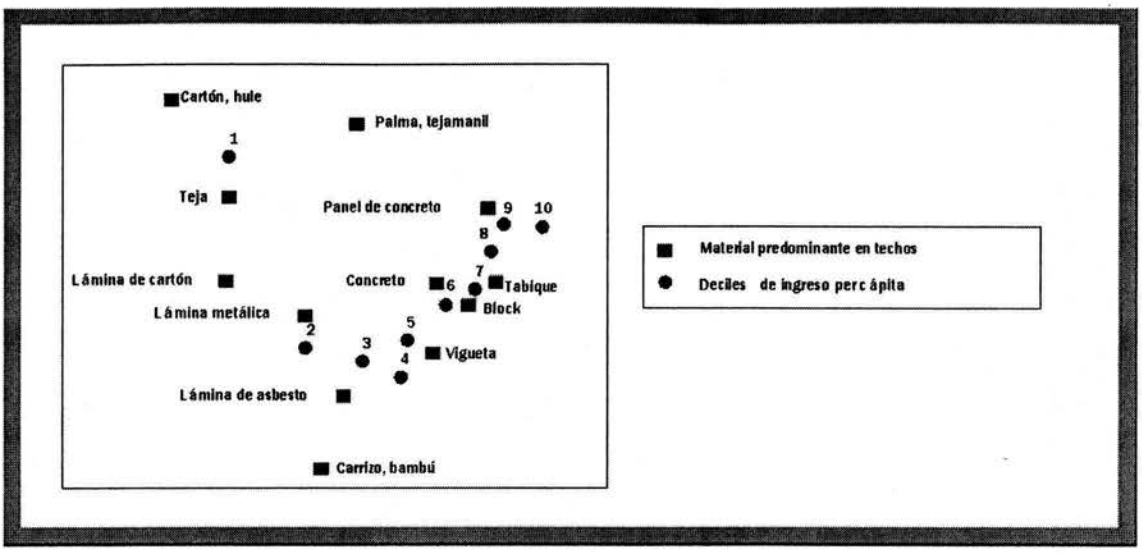
Apéndice II:

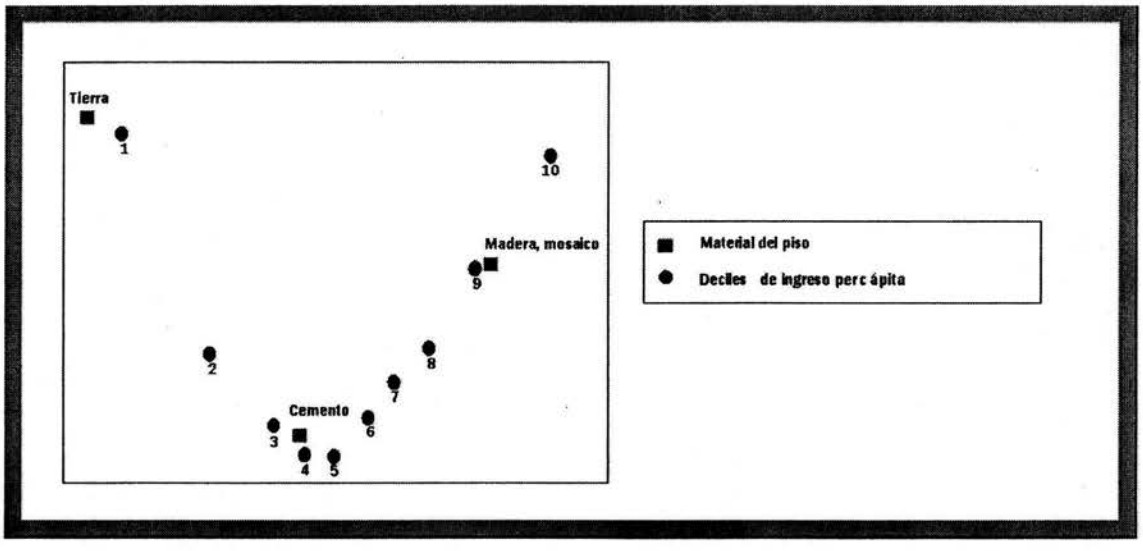
Análisis de Correspondencias Simples (deciles de ingreso vs. indicadores socioeconómicos y demográficos)

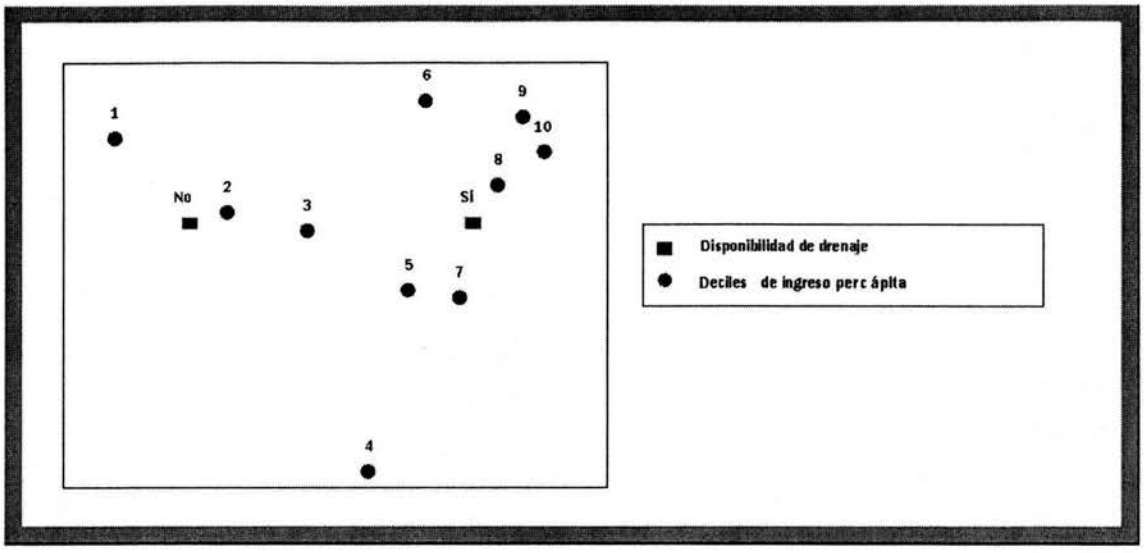
A continuación se presentan los análisis de correspondencias que se realizaron:

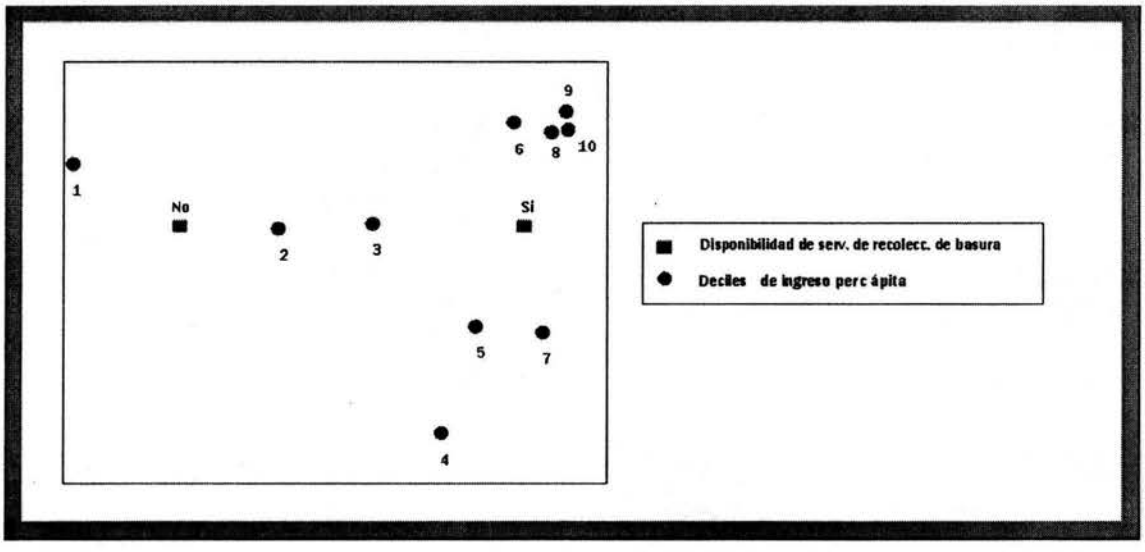


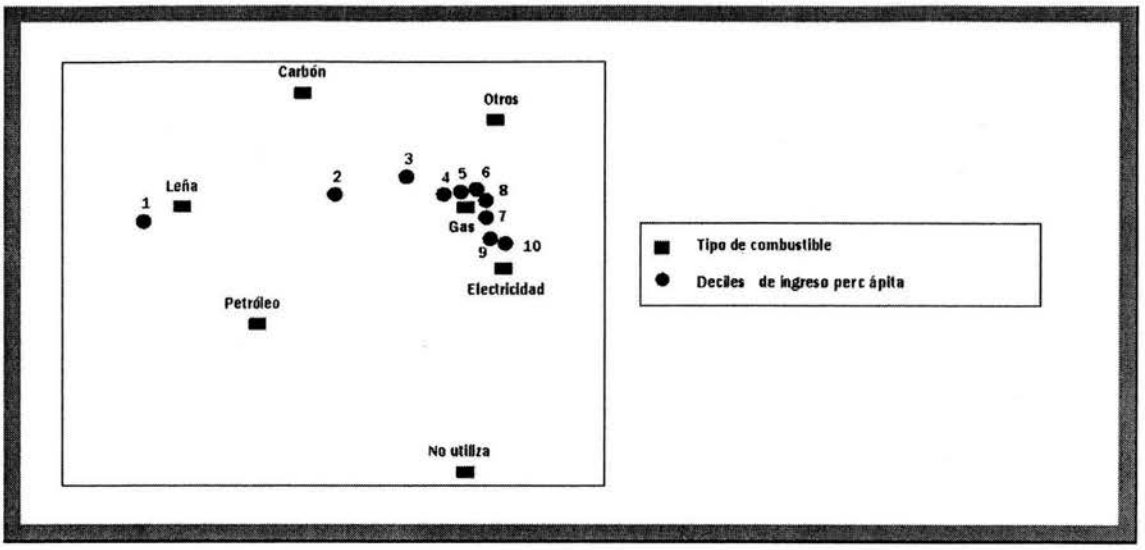


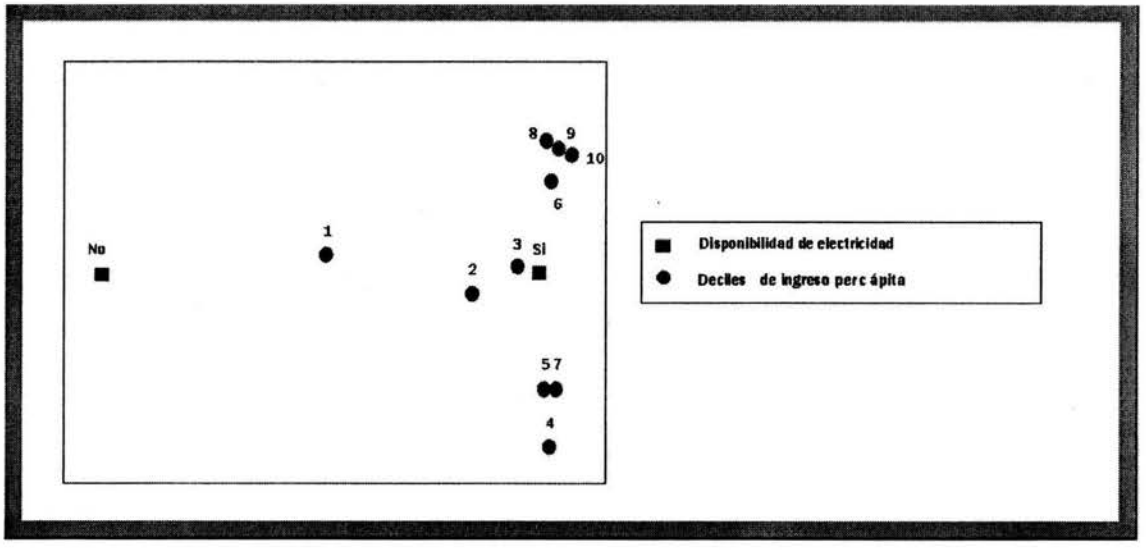

Cartón, hule
Palma, tejamanil
1
Teja
Panel de concreto
9
10
8
7
Lámina de cartón
Concreto
6
Tabique
Block
Lámina metálica
2
5
3
4
Vigueta
Lámina de asbesto
Carrizo, bambú
Material predominante en techos
Deciles de ingreso perc ápita

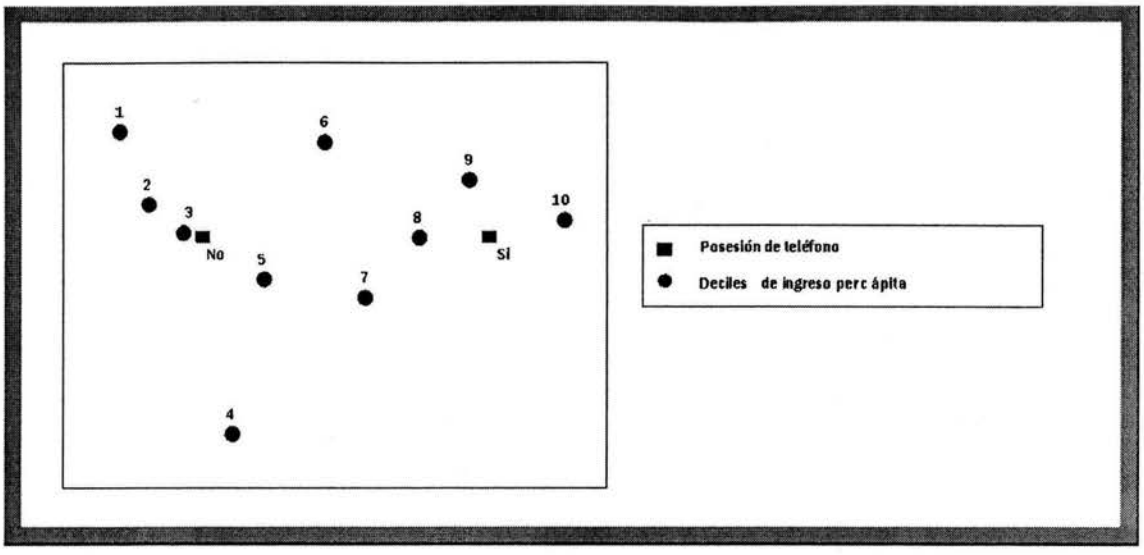

Tierra
1
10
Madera, mosaico
9
2
8
7
Cemento
6
3
4
5
Matedal del piso
Deciles de ingreso perc ápita

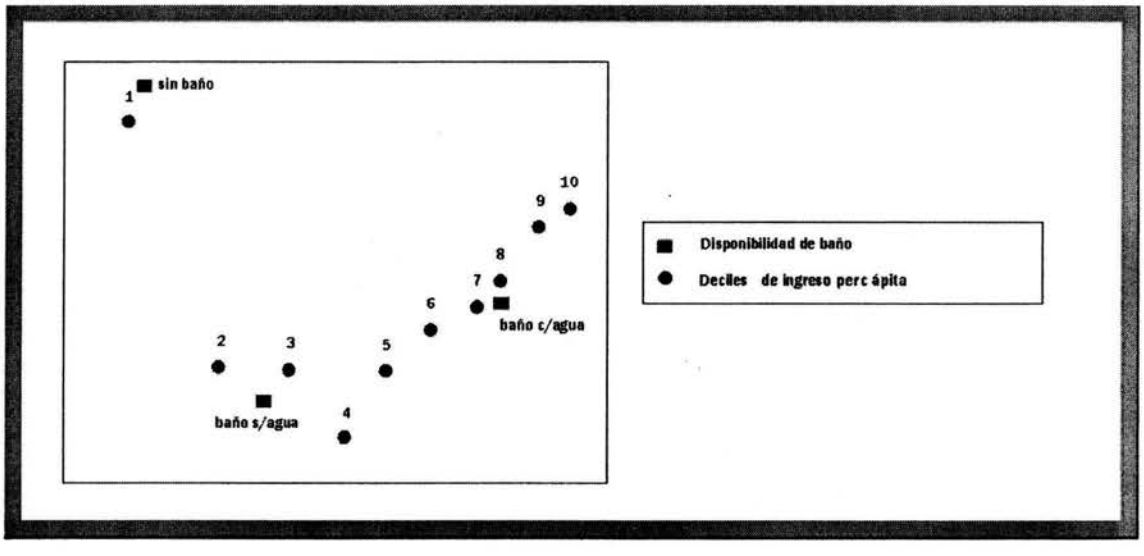

6
9
1
10
8
No
2
3
Sí
5
7
4
Disponibilidad de drenaje
Deciles de ingreso perc ápita

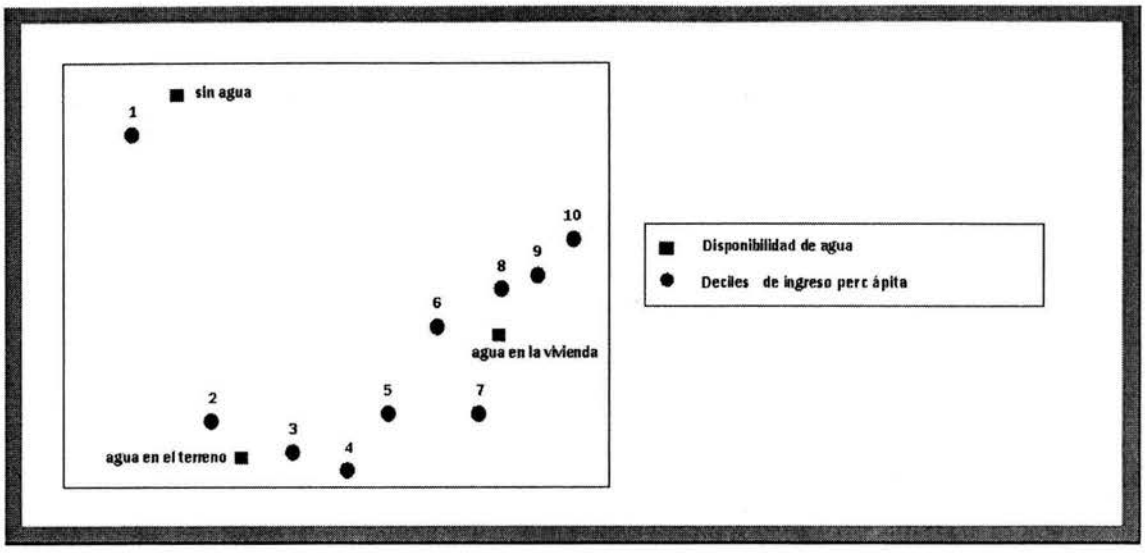

9
6
8
10
1
No
Sí
2
3
5
7
4
Disponibilidad de serv. de recolecc. de basura
Deciles de ingreso perc ápita

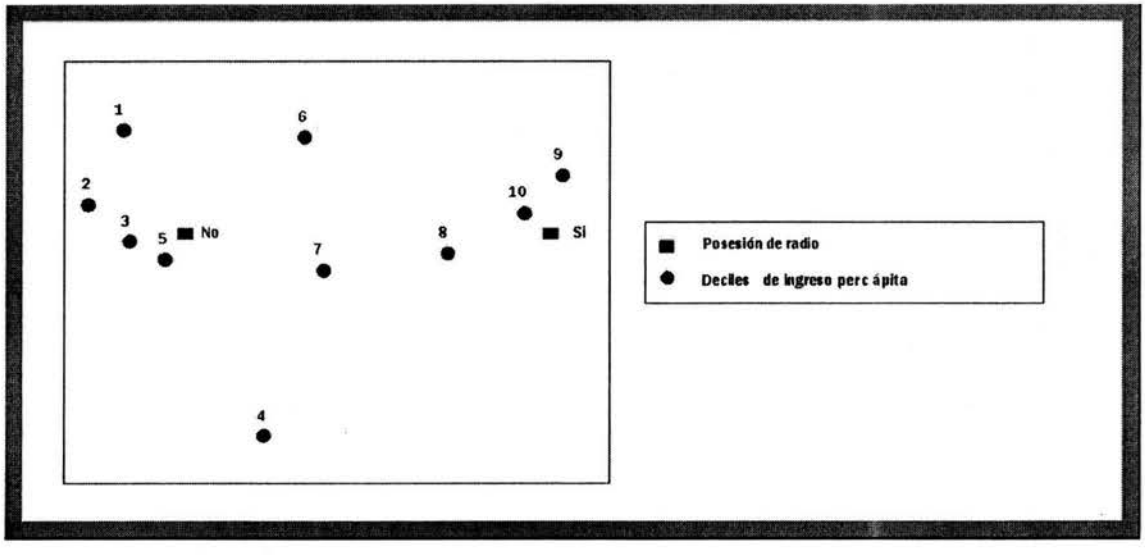

Carbón
Otros
Leña
Gas
Electricidad
Petróleo
No utiliza
1
2
3
4
5
6
7
8
9
10
Tipo de combustible
Deciles de ingreso perc ápita

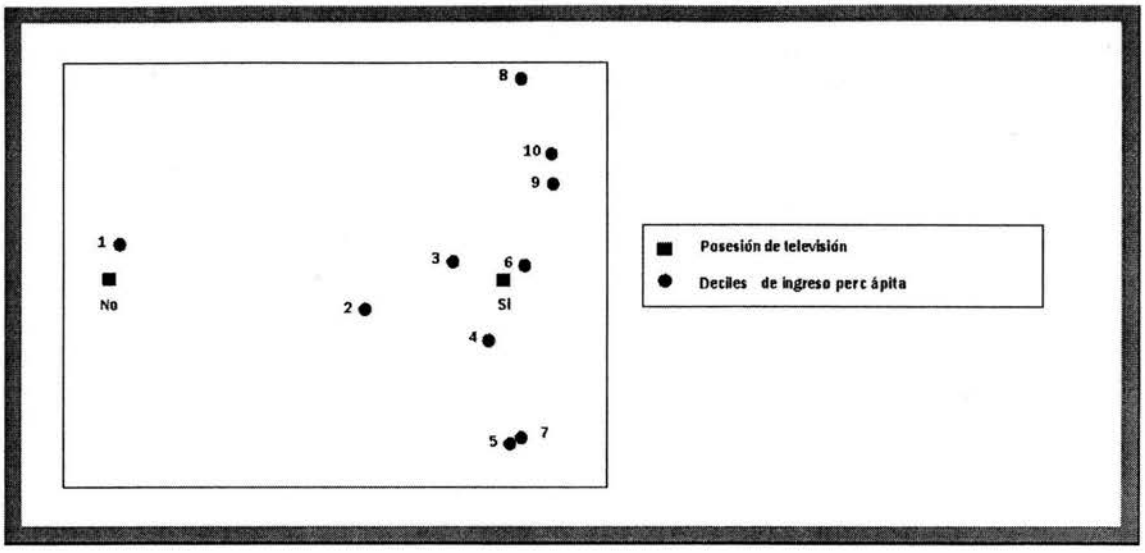

No
Si
1
2
3
4
5
6
7
8
9
10
Disponibilidad de electricidad
Deciles de ingreso perc ápita

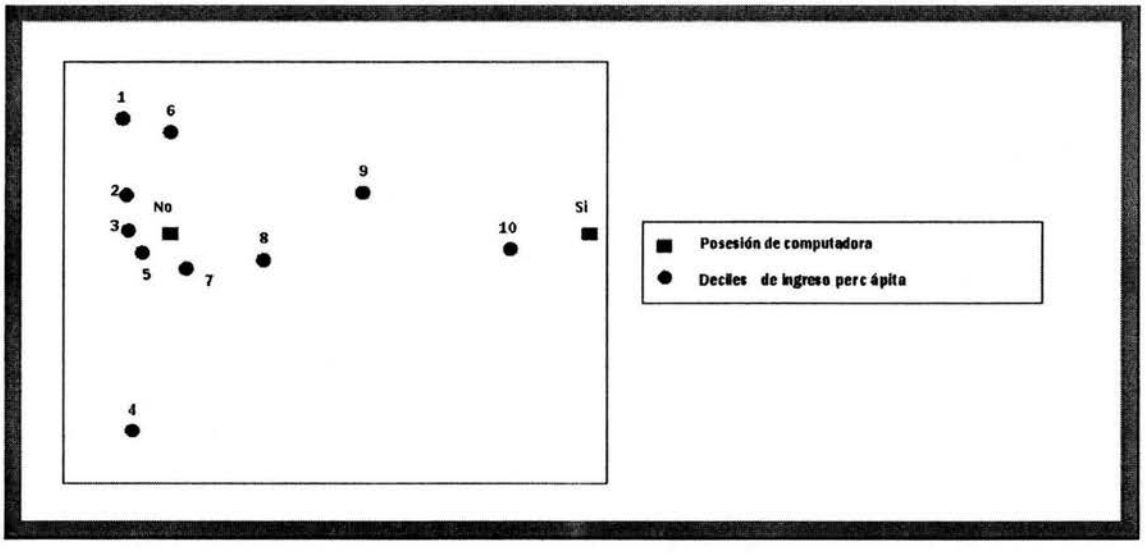

1
6
9
2
10
3
8
No
Sí
5
7
4
Posesión de teléfono
Deciles de ingreso perc ápita

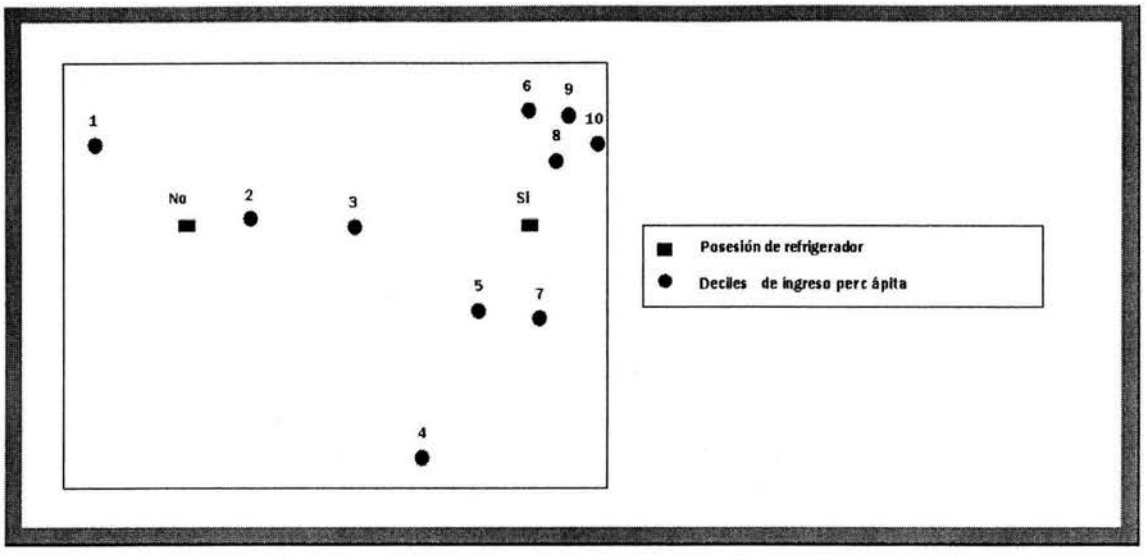

sin baño
1
10
9
8
7
6
baño c/agua
2
3
5
baño s/agua
4
Disponibilidad de baño
Deciles de ingreso perc ápita

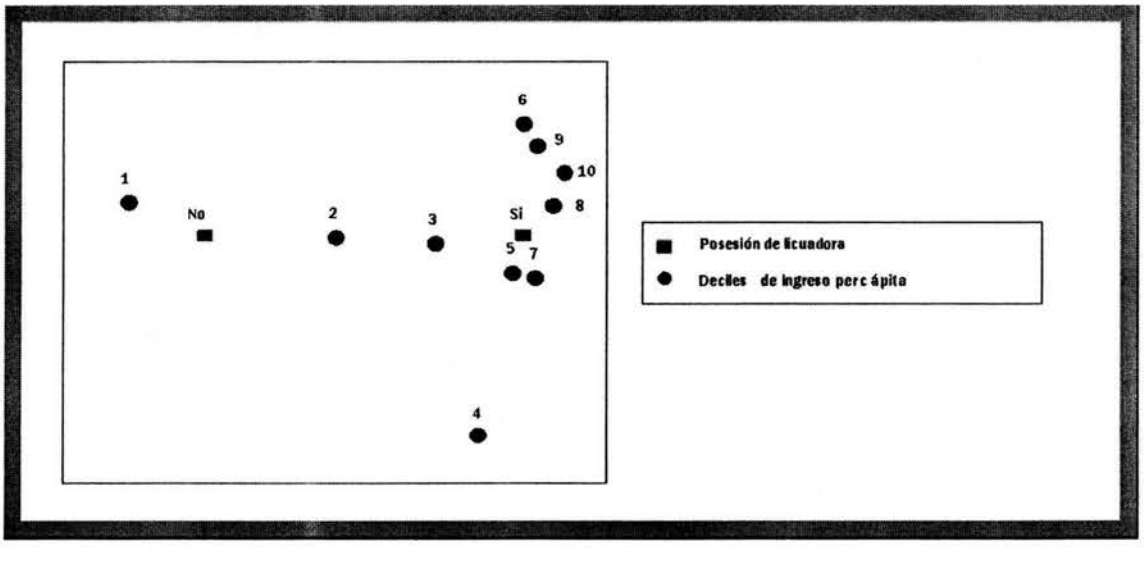

sin agua
1
10
9
8
6
agua en la vivienda
2
5
7
3
4
agua en el terreno
Disponibilidad de agua
Deciles de ingreso perc ápita

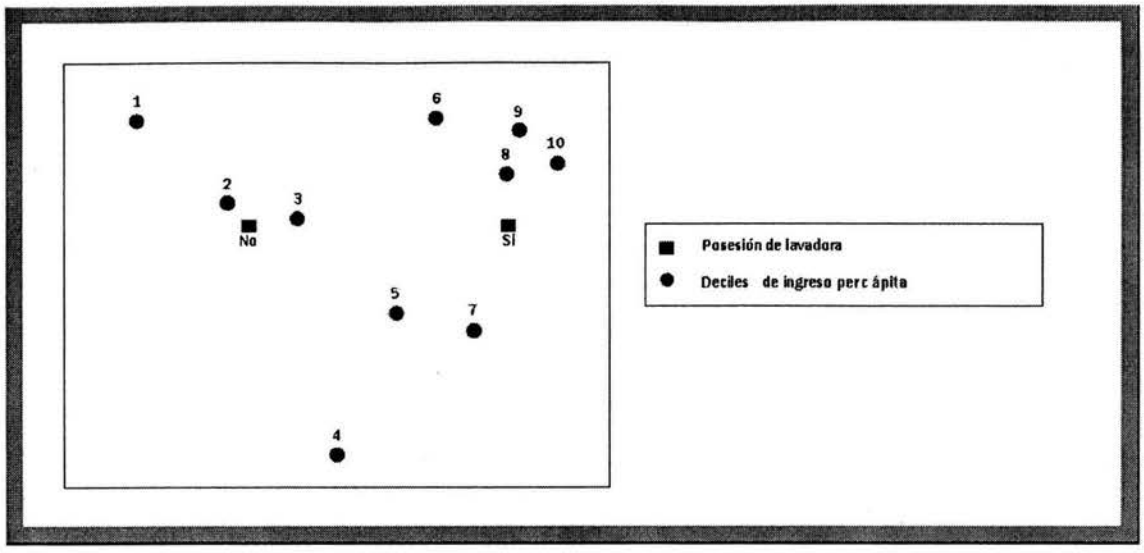

1
6
9
2
10
3
5
No
8
Si
7
4
Posesión de radio
Deciles de ingreso perc ápita

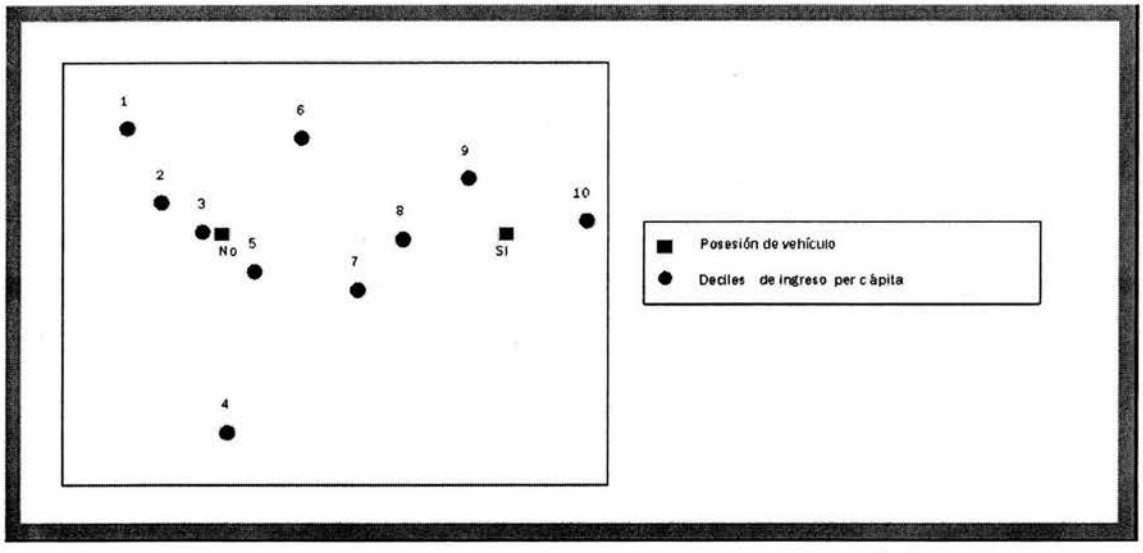

8
10
9
1
3
6
No
2
Si
4
5
7
Posesión de televisión
Deciles de ingreso perc ápita


1
6
9
2
No
Si
3
8
10
5
7
4
Posesión de computadora
Deciles de ingreso perc ápita


1
6
9
10
8
No
2
3
Si
5
7
4
Posesión de refrigerador
Deciles de ingreso perc ápita


1
6
9
10
8
No
2
3
Si
5
7
4
Posesión de licuadora
Deciles de ingreso perc ápita


1
6
9
10
8
2
3
No
Sí
5
7
4
Posesión de lavadora
Deciles de ingreso per cápita


1
6
9
2
10
3
8
No
5
Sí
7
4
Posesión de vehículo
Deciles de ingreso per cápita

Apéndice III:

Programa de generación de grupos por medio del procedimiento de "simulated annealing"

```
ki=100;
ntot=9986;
ni=[0 100 200 300 399 499 599 699 799 899 999 1098 1198 1298 1398 1498 1598 1698
1797 1897 1997 2097 2197 2297 2397 2497 2596 2696 2796 2896 2996 3096 3196 3295 3395
3495 3595 3695 3795 3895 3994 4094 4194 4294 4394 4494 4594 4693 4793 4893 4993 5093
5193 5293 5392 5492 5592 5692 5792 5892 5992 6091 6191 6291 6391 6491 6591 6691 6790
6890 6990 7090 7190 7290 7390 7490 7589 7689 7789 7889 7989 8089 8189 8288 8388 8488
8588 8688 8788 8888 8987 9087 9187 9287 9387 9487 9587 9686 9786 9886 9986];
nu=2;

for j=1:ki
    for i=(ni(j)+1):ni(j+1)
        ti{j}(i-ni(j),:)=gtot(i,:);
    end
end

for j=1:ki
    tpi{j}= ti{j}(1,:);
    for i=(ni(j)+2):ni(j+1)
        tpi{j}= tpi{j} + ti{j}((i-ni(j)),:);
    end
    tpi{j}= tpi{j}/(ni(j+1)-ni(j));
end

for j=1:ki
    SI{j}=(ti{j}(1,:)-tpi{j})'*(ti{j}(1,:)-tpi{j});
    for i=(ni(j)+2):ni(j+1)
        SI{j}=(ti{j}((i-ni(j)),:)-tpi{j})'*(ti{j}((i-ni(j)),:)-tpi{j})+SI{j};
    end
    SI{j}= SI{j}/(ni(j+1)-ni(j));
end
```

```
for j=1:ki
   SIinv{j}=inv(SI{j});
end

clear SI

for j=1:ki
   for i=(ni(j)+1):ni(j+1)
      for s=1:ki
         PPI(i,s)=(1+((1/((ni(s+1)-ni(s))+1))*(tpi{s}-ti{j}((i-ni(j)),:))
         *SIinv{s}*(tpi{s}-ti{j}((i-ni(j)),:))'))^(-((ni(s+1)-ni(s))
         -nu+(ki-1)+1)/2);
      end
      PFI(i,1)=j;
   end
end

for j=1:ki
   alfai=ni(j+1)-ni(j)-nu+1;
   zi=14;
   Psii=((alfai)/(alfai+nu))*SIinv{j};
   di=round((alfai+zi)/2);
   d2i=di/(alfai*pi);
   for i=1: ((zi/2)-1)
      di=di*(d2i-i) /(alfai*pi);
   end
   ci(j)=di*(det(Psii)^(1/2));
end

for j=1:ki
   for i=1:ntot
      PPI(i,j)=ci(j)*PPI(i,j);
   end
end

for i=1:ntot
   c2i(i)=PPI(i,1);
```

```
    for j=2:ki
        c2i(i)=c2i(i)+PPI(i,j);
    end
end

for j=1:ki
    for i=1:ntot
        PPI(i,j)= PPI(i,j) /c2i(i);
    end
end

for i=1:ntot
    QI(i)=max(PPI(i,:));
end

for i=1:ntot
    for j=1:ki
        if QI(i)==PPI(i,j) PFI(i,2)=j;
        end
    end
end

ui=0;
for j=1:ki
    for i=ni(j)+1:ni(j+1)
        ui=ui+(PPI(i,j)*log(PPI(i,j)));
    end
end

clear ti tpi SIinv alfai Psii di d2i ci c2i zi QI PPI PFI

crit=0;
tempii=1;

for i=1:23
    aux(i)=round((rand(1)*(100))+0.5);
end
```

```
clear aux it=1;
ut(it)=ui;
criter(it)=0;

while crit<10
   kii=ki-1;
   wii=round((rand(1)*(kii))+0.5);
   for j=1:kii+1
      if j<=wii nii(j)=ni(j);
      end
      if j>wii nii(j)=ni(j+1);
      end
   end

nu=2;
   for j=1:kii
      for i=(nii(j)+1):nii(j+1)
         tii{j}(i-nii(j),:)=gtot(i,:);
      end
   end

for j=1:kii tpii{j}= tii{j}(1,:);
      for i=(nii(j)+2):nii(j+1)
         tpii{j}= tpii{j} + tii{j}((i-nii(j)),:);
      end
      tpii{j}= tpii{j} /(nii(j+1)-nii(j));
   end

for j=1:kii
      SII{j}=(tii{j}(1,:)-tpii{j})'*(tii{j}(1,:)-tpii{j});
      for i=(nii(j)+2):nii(j+1)
         SII{j}=(tii{j}((i-nii(j)),:)-tpii{j})'*(tii{j}((i-nii(j)),:)-tpii{j})+SII{j};
      end
      SII{j}= SII{j}/(nii(j+1)-nii(j));
   end

for j=1:kii
      SIIinv{j}=inv(SII{j});
```

```
    end

clear SII

for j=1:kii
        for i=(nii(j)+1):nii(j+1)
            for s=1:kii
                PPII(i,s)=(1+((1/((nii(s+1)-nii(s))+1))*(tpii{s}-tii{j}
                ((i-nii(j)),:))*SIIinv{s}*(tpii{s}-tii{j}
                ((i-nii(j)),:))')^(-((nii(s+1)-nii(s))-nu+(kii-1)+1)/2);
            end
            PFII(i,1)=j;
        end
    end

for j=1:kii
        alfaii=nii(j+1)-nii(j)-nu+1;
        zii=14;
        Psiii=((alfaii)/(alfaii+nu))*SIIinv{j};
        dii=round((alfaii+zii)/2);
        d2ii=dii/(alfaii*pi);
        for i=1: ((zii/2)-1)
            dii=dii*(d2ii-i) /(alfaii*pi);
        end
        cii(j)=dii*(det(Psiii)^(1/2));
    end

for j=1:kii
        for i=1:ntot
            PPII(i,j)=cii(j)*PPII(i,j);
        end
    end

for i=1:ntot
        c2ii(i)=PPII(i,1);
        for j=2:kii
            c2ii(i)=c2ii(i)+PPII(i,j);
        end
```

```
    end

for j=1:kii
        for i=1:ntot
            PPII(i,j)= PPII(i,j) /c2ii(i);
        end
    end

for i=1:ntot
        QII(i)=max(PPII(i,:));
    end

for i=1:ntot
        for j=1:kii
            if QII(i)==PPII(i,j) PFII(i,2)=j;
            end
        end
    end

uii=0;
    for j=1:kii
        for i=nii(j)+1:nii(j+1)
            uii=uii+(PPII(i,j)*log(PPII(i,j)));
        end
    end

if uii>ui
        clear ki ni ui
        ki=kii;
        ni=nii;
        ui=uii;
        crit=0;
        tempii = tempii *1.2;
    elseif uii<=ui
        rvii=rand(1);
        valii=exp((-(ui-uii))/ tempii);
        if valii<rvii
            clear ki ni ui
```

```
            ki=kii;
            ni=nii;
            ui=uii;
            crit=0;
        ..tempii = tempii *1.2;
        elseif valii>=rvii
            crit=crit+1 ;
        end
    end

if crit<10
        clear PPII PFII
    end

clear tii tpii SIIinv alfaii Psiii dii d2ii cii c2ii zii wii QII kii nii uii
    it=it+1;
    ut(it)=ui;
    criter(it)=crit;
    clasif{it}=ni;

 end
```

Apéndice IV:

Programa de clasificación bayesiana

```
k=4;

ntot=9962;

n=[0 2882 3638 5812 9962];

nu=2;

    for j=1:k
       for i=(n(j)+1):n(j+1)
          g{j}(i-n(j),:)=gtot(i,:);
       end
    end

    for j=1:k
       gs{j}= g{j}(1,:);
       for i=2:(n(j+1)-n(j))
          gs{j}= gs{j} + g{j}(i,:);
       end
    end

    for j=1:k
       gns{j}=gs{j}-gs{j};
       for i=1:k
          if i =j
             gns{j}= gns{j}+ gs{i};
          end
       end
       gp{j}=gs{j}/(n(j+1)-n(j));
       gnp{j}=gns{j}/(ntot-(n(j+1)-n(j)));
    end
```

```
for j=1:k
   V{j}=(g{j}(1,:)-gp{j})'*(g{j}(1,:)-gp{j});
   for i=2:(n(j+1)-n(j))
      V{j}=(g{j}(i,:)-gp{j})'*(g{j}(i,:)-gp{j})+V{j};
   end
end

VT=V{1};
for j=2:k
   VT=V{j}+VT;
end

Vinv=inv(VT/ntot);

for j=1:k
   for s=1:(k-1)
      for i=(n(j)+1):n(j+1)
         t{j}((i-n(j)),s)=(gp{s}-gnp{s})*Vinv*gtot(i,:)';
      end
   end
end

for j=1:k
   tp{j}= t{j}(1,:);
   for i=(n(j)+2):n(j+1)
      tp{j}= tp{j} + t{j}((i-n(j)),:);
   end
   tp{j}= tp{j} /(n(j+1)-n(j));
end

for j=1:k
   S{j}=(t{j}(1,:)-tp{j})'*(t{j}(1,:)-tp{j});
   for i=(n(j)+2):n(j+1)
      S{j}=(t{j}((i-n(j)),:)-tp{j})'*(t{j}((i-n(j)),:)-tp{j})+S{j};
   end
   S{j}= S{j}/(n(j+1)-n(j));
end
```

```
    for j=1:k
       Sinv{j}=inv(S{j});
    end

    for j=1:k
       for i=(n(j)+1):n(j+1)
          for s=1:k
             P(i,s)=(1+((1/((n(s+1)-n(s))+1))*(tp{s}-t{j}((i-n(j)),:))
             *Sinv{s}*(tp{s}-t{j}((i-n(j)),:))'))^(-((n(s+1)-n(s))-nu+(k-1)+1)/2);
          end
          PF(i,1)=j;
       end
    end

    for j=1:k
       alfa=n(j+1)-n(j)-nu+1;
       z=k-1;
      Psi=((alfa)/(alfa+nu))*Sinv{j};
       d=round((alfa+z)/2);
       d2=d;
       for i=1: (round(z/2)-1)
          d=d*(d2-i);
       end
       c(j)=(d/((alfa*pi)^(z/2)))*(det(Psi)^(1/2));
    end

    for j=1:k
       for i=1:ntot
          P(i,j)=c(j)*P(i,j);
       end
    end

for j=1:k
    for i=1:ntot
       P(i,j)=((n(j+1)-n(j))/ntot)*P(i,j);
    end
 end
```

```
for i=1:ntot
    c2(i)=P(i,1);
    for j=2:k
       c2(i)=c2(i)+P(i,j);
    end
 end

 for j=1:k
    for i=1:ntot
       P(i,j)= P(i,j) /c2(i);
    end
 end

 for i=1:ntot
    Q(i)=max(P(i,:));
 end

 for i=1:ntot
    for j=1:k
       if Q(i)==P(i,j) PF(i,2)=j;
       end
    end
 end
```