SECRETARÍA DE DESARROLLO SOCIAL | SEDESOL

serie: documentos de investigación

LA FOCALIZACIÓN COMO ESTRATEGIA DE POLÍTICA PÚBLICA

Daniel Hernández F.
Mónica Orozco C.
Sirenia Vázquez B.

Noviembre 2005 25

SECRETARÍA DE DESARROLLO SOCIAL | SEDESOL

serie: documentos de investigación

LA FOCALIZACIÓN COMO ESTRATEGIA DE POLÍTICA PÚBLICA

Daniel Hernández F.*
Mónica Orozco C.
Sirenia Vázquez B.

*Los autores trabajan en la Secretaría de Desarrollo Social.
Daniel Hernández F. es Coordinador de Asesores de la Secretaria de Desarrollo Social; Mónica Orozco C. es Directora General de Análisis y Prospectiva; Sirenia Vázquez B. es Subdirectora de Información Estadística.
Las opiniones de los autores no reflejan necesariamente la posición oficial de la institución.

Noviembre 2005 25

2005
Secretaría de Desarrollo Social

"La focalización como estrategia de política pública"

Serie: *Documentos de Investigación,* 25

ISBN: 968-838-619-7

Dr. Gonzalo Hernández Licona
Coordinador de la serie

Emiliano Pérez Cruz
Coordinación editorial

Martha González Serrano
Formación editorial

Paseo de la Reforma 116
Col. Juárez, C.P. 06600
México, D.F.

Impreso en México / *Printed in Mexico*

Contenido

Introducción

Cuando se analizan los resultados de las políticas públicas, muchas veces se concluye que éstas no han favorecido a los pobres, o no con la eficacia esperada. De ahí el esfuerzo por lograr un mejor uso de los recursos disponibles mediante estrategias de focalización, que consisten en dirigir las acciones a una población o territorio definidos, para concentrar la atención sobre un determinado problema o necesidad. Esta orientación no es homogénea, sino que considera las peculiaridades de las poblaciones y las regiones, para desarrollar mecanismos adecuados que correspondan al objetivo que se busca.

El propósito de la focalización es asegurar que los beneficios de las acciones lleguen a las familias que más requieren las intervenciones públicas. En el caso de la política social, éstas son las familias más pobres. Se trata de lograr un mayor impacto per cápita que el que podría derivarse de una política general que se aplica por igual a toda la población. Es una orientación que busca propiciar la eficiencia en la gestión de los recursos.

Pero la focalización busca más que la sola eficiencia de los esfuerzos y los recursos que se aplican, ya que encierra también un principio de justicia: ante recursos necesariamente escasos para atender a todos o a todas las necesidades, tan importante es asegurar que se beneficien quienes más los necesitan, como no destinar recursos a quienes no se encuentran en una situación apremiante. La focalización es una forma de promover la equidad, por lo que su ausencia puede, incluso, ampliar las brechas de injusticia y aumentar la inequidad.

La focalización debe realizarse con objetividad, transparencia y sin discrecionalidad alguna. En el pasado, la falta de objetividad y cierto nivel de discrecionalidad han sido argumentados como elementos que hacían poco deseable un esquema de trabajo focalizado. No obstante, ahora se cuenta con herramientas técnicas y procesos que, por el contrario, favorecen la credibilidad en la imparcialidad y equidad de las acciones de política social.

Pero la focalización también se concibe como un instrumento para disminuir el clientelismo. En muchos casos, mecanismos más tradicionales de operación de políticas públicas no han llegado a los más pobres, por cuanto han tendido a satisfacer las demandas de los grupos que tienen menores necesidades, pero cuya ubicación territorial tiene más fáciles accesos, cuentan con organización social y política, o disponen de mayor información para acceder a los programas y proyectos sociales. Por el contrario,

los hogares en condición de pobreza enfrentan barreras culturales y procedimientos burocráticos o requisitos que no pueden satisfacer, carecen de acceso a la información sobre las acciones de la política social, tienen escaso peso político que les hace difícil defender adecuadamente sus derechos, y afrontan costos de transacción (como el costo de transportarse al lugar donde se solicitan los apoyos y el tiempo que invierten) que dificulta la demanda, incluso, de los servicios gratuitos.

Sin embargo, la focalización tiene costos. La relativa falta de voz de los pobres frente a otros grupos de la sociedad puede introducir un costo político para el consenso y el apoyo sostenido a la focalización. Las personas pueden distorsionar la información para ser identificados como posibles beneficiarios de las acciones públicas, e incluso pueden presentarse incentivos perversos para permanecer en el grupo de beneficiarios. Adicionalmente, ser identificado en un programa focalizado puede ser considerado un estigma. Y es indispensable considerar que la identificación de los beneficiarios conlleva un costo administrativo que debe analizarse con medidas de costo-efectividad. Muchos de estos temas, sin embargo, son atendibles; se han construido mecanismos de respuesta amplia, estos deben considerarse contra la alternativa de dirigir recursos a quienes no los necesitan, que implica un costo ético y administrativo.

La focalización perfecta no existe, ni en teoría, ni en la práctica. Lo que se busca es un marco eficiente y justo que garantice el máximo beneficio a los grupos más desfavorecidos. Todos estos son aspectos fundamentales para ser cuidados en el proceso de diseño de acciones focalizadas. De hecho, como en todo diseño de políticas, es necesario tomar en consideración diversos elementos para definir esta orientación: la viabilidad, el costo, los incentivos que genera y la efectividad.

Aunque la focalización aparece como un concepto sencillo, en la práctica se necesita utilizar complejas herramientas técnicas para realizarla adecuadamente. Focalizar representa un aspecto esencial de modernización de la política social, para lo cual es necesario contar con información estratégica. Dentro del mismo marco, la información es un instrumento para impulsar la integración de las políticas sociales entre sí, a la vez que se busca asegurar que no se dispersen, atomicen, ni se dupliquen los beneficios de las acciones. También resulta ser una herramienta clave para la generación de evidencia y la evaluación de los resultados e impacto de los programas que deben ser aplicados en el proceso de mejora continua del diseño e implementación de las acciones.

El objetivo de este trabajo es documentar y analizar algunas de las herramientas que se utilizan en la política social para acercar los programas a la población que vive en condiciones de pobreza. En el primer apartado se explican los principios generales

y ventajas de los apoyos focalizados, y se describen tres ejemplos de programas sociales coordinados por la Secretaría de Desarrollo Social, que se caracterizan por ser focalizados a distintos niveles. En el segundo apartado se detalla la metodología estadística y econométrica que se utiliza en los procesos de focalización en México y se comparan los resultados con aproximaciones estadísticas alternativas para la identificación de hogares susceptibles de recibir apoyos para la superación de la pobreza. Finalmente, se presentan algunos de los posibles retos para mejorar las herramientas de planeación, seguimiento y evaluación relacionadas con la implementación de programas focalizados en México.

1. Apoyos focalizados

1.1 Apoyos dirigidos

En la última década, tanto en México como en muchos otros países, se ha impulsado la implantación de políticas focalizadas dentro del ámbito del desarrollo social. En la actualidad podemos decir que la política social está consolidando los esquemas focalizados, con el fin de dirigir la mayor parte de los recursos hacia la población que enfrenta mayores niveles de vulnerabilidad.

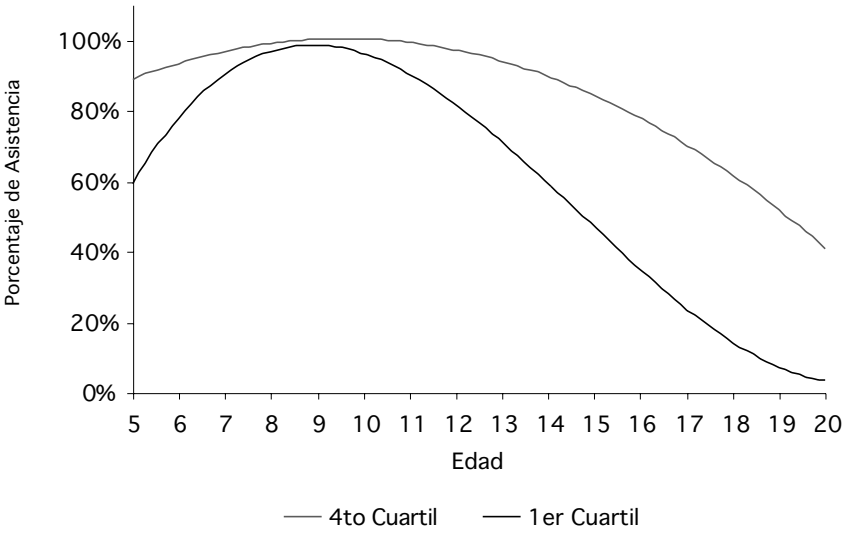
De forma simplificada, puede decirse que los apoyos focalizados son aquellos que se dirigen hacia grupos de población que presentan características específicas o que se encuentran habitando en zonas delimitadas territorialmente.

Una de las características más importantes de los apoyos focalizados es que parten de la premisa de que no existe un acceso equitativo a los bienes o servicios para toda la población, y que sólo una direccionalidad intencionada ayuda a que quienes menos tienen puedan superar los obstáculos para el aprovechamiento de los apoyos.

Para dar un ejemplo, analicemos brevemente el caso de la educación. En México, la asistencia escolar durante los primeros nueve grados es gratuita. El objetivo es que toda la población satisfaga el derecho a la educación básica facilitando el acceso a la misma de manera equitativa. Esta política se lleva a cabo a través de apoyos que podemos denominar “por el lado de la oferta”, es decir, son apoyos dirigidos a disminuir total o parcialmente el costo de un bien o servicio, en este caso los servicios educativos, bajo la premisa de que todo aquel que lo desee podrá acceder al mismo sin restricción alguna dado que son de carácter gratuito.

Sin embargo, la realidad indica que existen ciertas características de algunos grupos de la población que obstaculizan el aprovechamiento óptimo de los apoyos aplicados en forma generalizada. Las condiciones de rezago y pobreza en las que vive un segmento importante de la población, provocan muchas veces que la apropiación de los beneficios de la educación gratuita se vea limitada al no disponer de mecanismos que les permitan acceder a ella, debido principalmente a la existencia de otro tipo de costos “asociados” que se requiere cubrir. En la gráfica 1, se evidencia que la asistencia escolar de los niños de los hogares de bajos ingresos es menor que la de los niños de grupos de mayores recursos económicos.

Gráfica 1
Asistencia escolar por cuartil de ingreso



Fuente: Cálculos propios con el XII Censo General de Población y Vivienda, 2000.

Comúnmente sucede que quienes se ubican en los estratos de mayor pobreza resultan menos beneficiados por los apoyos no focalizados que los ciudadanos de mayores recursos. Esto tiene que ver con el hecho de que para poder hacer uso pleno de los apoyos de carácter general se requiere contar con otros recursos que no necesariamente están disponibles a toda la población. Los ejemplos son múltiples, pero entre los más importantes está el hecho de que las familias de menos recursos no cuentan con los ingresos suficientes para cubrir los costos del transporte de sus hijos a la escuela u otros gastos (útiles escolares, almuerzos, uniformes), o en muchas ocasiones requieren de la participación laboral de los hijos para contribuir al sostenimiento del hogar (costo de oportunidad del tiempo de los niños).

El caso de la educación superior y media superior gratuita es también un ejemplo de un apoyo cuya intención es dar acceso a todos los grupos de población, pero que en la práctica es aprovechado en mayor medida por quienes cuentan con mejores condiciones socioeconómicas (Scott, 2004). Este tipo de apoyos resultan regresivos, esto es, benefician principalmente a los estratos de población de más elevados ingresos, en comparación con los beneficios que reciben los hogares de menos recursos (Cuadro

1).[1] Ante estas imperfecciones, una posible alternativa es la focalización de apoyos adicionales, ya sea por el "lado de la oferta" o por el "lado de la demanda", a la población que enfrenta mayor vulnerabilidad.

Cuadro 1
Distribución del gasto público en Educación según deciles de ingreso, 2002

Deciles	Educación Terciaria
1	0.0%
2	0.7%
3	2.4%
4	4.5%
5	6.7%
6	8.2%
7	11.1%
8	21.9%
9	22.1%
10	22.3%
Urbano	97.7%
Rural	2.3%

Fuente: Scott, J., (2004). Información obtenida con datos de la ENIGH (2002). Deciles de población ordenados por gasto per cápita.

Un ejemplo de la aplicación de apoyos por el "lado de la oferta", lo constituyen los recursos para el mejoramiento de la calidad de los servicios y disponibilidad de los mismos únicamente en determinadas zonas territorialmente definidas que, por sus características de aislamiento, concentración de hogares en pobreza o marginación, sean susceptibles de recibir más apoyos.

Por el "lado de la demanda", la focalización de apoyos implica la segmentación de la población con la intención de que sólo determinados grupos, caracterizados por sus niveles de pobreza o vulnerabilidad, reciban los apoyos. En este caso, en vez de dirigir los recursos a mejorar la infraestructura o el abastecimiento de los servicios que se proporcionan a la población, se entregan de manera directa a las personas con el fin de acrecentar el uso que éstas hacen de los bienes o servicios. Un ejemplo claro de este tipo de apoyos son las becas que reciben los hogares con menos recursos para que los hijos destinen su tiempo a los estudios y puedan cubrir los gastos asociados con la asistencia a la escuela (gratuita).

Una de las razones fundamentales por las que resulta de especial importancia la aplicación de políticas focalizadas, es que por más eficientes que sean los sistemas de apoyos generales, la proporción de los recursos que llega a quienes viven en con-

[1] Ver Banco Mundial (2004).

diciones de pobreza es escasa en comparación con el gasto total que se realiza. En tanto que la población que se busca apoyar represente solamente una proporción de la población total, la utilización de políticas focalizadas es mucho más eficiente frente a una aplicación de recursos de carácter general, que implica un determinado nivel de dispendio de los mismos. Pensar que un gasto generalizado es "universal" es inexacto, porque no siempre se cumple necesariamente que los supuestos beneficios lleguen a todos, e inclusive no es infrecuente que precisamente se apoye proporcionalmente más a quienes no lo necesitan.

Esto último es relevante porque se llega a argumentar la idea de que los apoyos generales pueden evitar un uso poco eficiente si las personas deciden no utilizar el apoyo ofrecido (se autoseleccionan para no utilizarlo si no lo requieren). Existen, por supuesto, otros mecanismos para "impulsar" la autoselección: apoyos en especie con productos específicos, suponiendo que las personas de mayores ingresos no retirarán (lo que puede llevar a otorgar apoyos con bienes de "segunda clase"); o la existencia de "colas" para retirar los apoyos, suponiendo que los costos de tiempo de espera son más elevados y por lo tanto menos aceptables para los grupos en mejores condiciones sociales.

Una de las opciones más utilizadas para medir la eficiencia de una política focalizada en comparación con una política de corte general, son los análisis de costo-beneficio. También es posible realizar comparaciones de eficiencia entre políticas focalizadas, donde las diferencias son atribuibles tanto al propio mecanismo de focalización, como al nivel de desagregación al cual se realiza ésta. Skoufias, Davis y Behrman (2000) muestran cómo los apoyos generalizados implican tasas de "fuga" de recursos sustantivas y una menor eficiencia en la atención a los más pobres, en comparación con métodos focalizados geográfica o individualmente.

Un estudio realizado por Coady (2003) indica que los mecanismos de focalización son por lo general más efectivos en países con mayores niveles de desigualdad. México es un país con elevados niveles de desigualdad, razón por la cual el concepto de focalización es importante en el diseño de políticas sociales.

La revisión de 111 mecanismos de focalización aplicados recientemente en el mundo indica que los más efectivos se caracterizan por involucrar más de un criterio de focalización. Por ejemplo, la utilización de focalización geográfica, esto es, dirigir los apoyos a determinadas zonas delimitadas territorialmente, y posteriormente de focalización individual, es decir, identificación de familias o personas con determinadas condiciones de vulnerabilidad, dentro de las zonas geográficas ya establecidas, permite refinar los procedimientos de identificación. Coady demuestra que las intervenciones

que utilizan mecanismos de focalización a través de "pruebas de medios",[2] es decir, de mecanismos que miden los medios con que las personas cuentan, están asociadas con porcentajes de beneficios mayores para los hogares de menores recursos. Los medios que se utilizan pueden ser, por ejemplo, ingresos, años de escolaridad o tipo de empleo, que estásn asociados con la capacidad de las personas para funcionar.

Los resultados obtenidos por Coady (2003) y Skoufias, et. al. (2000) reflejan que se cumple la premisa fundamental sobre la cual se sustentan las estrategias focalizadas: que las acciones se dirigen hacia los segmentos de población de menores recursos con el fin de mejorar sus condiciones de vida y de disminuir la desigualdad respecto de otros sectores de la población que cuentan con mayores recursos.

1.2 Microdatos y Sistemas de Información Geo-referenciada

Actualmente, la mayoría de los mecanismos de focalización en las políticas sociales se valen de información estadística disponible a distintos niveles de agregación para identificar a los posibles beneficiarios.

Sin embargo, pueden señalarse otros mecanismos de identificación: el basado en la opinión de un funcionario, que partiendo de una investigación propia (estructurada o no), determina la condición de susceptibilidad de recibir los apoyos de alguna persona, familia o grupo; otro tipo es aquel que aprovecha la experiencia de una comunidad, para determinar quiénes deben recibir los apoyos ofrecidos. El riesgo de estos dos esquemas, es que se incurre en cierto nivel de discrecionalidad de los individuos que toman las decisiones, pudiendo llegarse al extremo de favoritismo en la identificación de los posibles beneficiarios de los apoyos por razones económicas o políticas. Adicionalmente, los criterios varían de un individuo a otro, o de una comunidad a otra, por lo que la regla de decisión que se aplica es distinta dependiendo de quién lo sugiere y del contexto en el que se ubica, perdiéndose así la visión de conjunto con la que se puede dar prioridad a los más necesitados de un país o región.

Contar con datos estratégicos constituye una parte indispensable para el diagnóstico de la población, la planeación, seguimiento y evaluación de los programas sociales y las estrategias para la superación de la pobreza. La tarea de integración de sistemas de información implica, entre otras acciones, la recolección y/o sistematización de diversas fuentes de datos, la aplicación de definiciones metodológicas, la utilización de herramientas estadísticas y la incorporación de tecnología para el manejo

[2] Traducción de los autores del concepto en inglés: "proxy means testing".

de datos. Los mecanismos de focalización que a continuación se discuten, serían inviables sin los avances en el sistema estadístico nacional y la disponibilidad de poderosas tecnologías para el procesamiento de datos actualmente existentes.

Comúnmente, los microdatos corresponden a personas o unidades de análisis básicas y constituyen la unidad última de desagregación para la cual hay información disponible. A partir de este tipo de información se construyen datos a distintos niveles de agregación que permiten el análisis de eventos que ocurren a nivel de grupos, como por ejemplo: hogares estructurados por una o más personas, que habitan en comunidades, que integran municipios, que a su vez conforman estados y estos a su vez países.

Los niveles de agregación para los cuales se puede contar con información dependen del tipo de estructura de anidamiento de los datos primarios o microdatos y del nivel de detalle utilizado en la recolección de la información. En general, entre más desagregado sea el nivel disponibilidad de la información se puede contar con información más detallada. Sin embargo, dependiendo de la varianza de los datos hacia el interior de cada grupo de anidamiento y de los fines que se persigan, la información puede ser útil a niveles de agregación mayores.

Tomemos por ejemplo los datos del Censo General de Población y Vivienda y el Índice de Marginación que el Consejo Nacional de Población (Conapo) calcula con ellos. Estos datos se recaban a nivel de los hogares, registrando información del hogar en su conjunto y de determinadas características de cada uno de sus integrantes. Esta es una información muy desagregada.

En la práctica, para efectos de presentación, generalmente se agregan los datos de los hogares que provienen de estas fuentes por entidad federativa. Utilizando diversas variables como el porcentaje de hogares con piso de tierra o la proporción de personas mayores de 15 años analfabetas se calcula un Índice de Marginación.[3] Como herramienta de planeación, reconocer a las entidades federativas con más rezagos es útil, pero para alcanzar mayor precisión en el direccionamiento de las acciones, resulta muchas veces más conveniente precisar este Índice a nivel de los municipios (espacios territoriales administrativos más específicos) o incluso a nivel de las localidades (espacios aún más pequeños por lo general, salvo en el caso de algunas ciudades). El Índice de Marginación, dado su mecanismo de construcción, se puede calcular a cualquiera de estos tres niveles. De hecho, esto revela muchas veces importantes heterogeneidades al interior de los territorios, de manera que en estados con muy baja marginación

[3] El índice de marginación es una medida que permite diferenciar entidades federativas y municipios según el impacto de las características socioeconómicas relacionadas con la población. La información se toma del Censo General de Población y Vivienda y las características que se incluyen son: educación, vivienda, densidad de población y empleo. Para una descripción más detallada sobre la construcción del índice ver Índices de Marginación CONAPO (1995 y 2000).

se pueden encontrar municipios marginados, y a su vez en municipios bajamente marginados, se ubican localidades altamente marginadas.

Con esto, se quiere ejemplificar cómo una misma fuente de datos puede ser analizada de distintas maneras dependiendo del grado de desagregación de los datos. Se puede establecer que por lo general, la información con mayor nivel de desagregación es más específica. Sin embargo, la selección de los indicadores que se utilizan para focalizar acciones depende directamente del objetivo que se quiere lograr.[4]

Mediante datos estadísticos, es posible construir diversos indicadores técnicamente robustos. En particular, la operación de programas dirigidos hacia la población de menores recursos requiere contar con mecanismos de planeación basados en indicadores de la condición de pobreza de las personas o de los hogares.

Pero conocer el número o proporción de personas o comunidades en determinada condición no es del todo suficiente. La aplicación de los apoyos tiene un componente físico-espacial: se necesitan entregar en un determinado lugar. Esto es, se necesita pasar de la relativa abstracción de los datos estadísticos sobre la magnitud o prevalencia de la pobreza a la identificación de los espacios precisos en que se ubican las personas o grupos que se desean apoyar.

Hasta hace relativamente poco tiempo no existía la tecnología para vincular a través de una sola herramienta información estadística con información territorial, pero en la actualidad se dispone de Sistemas de Información Geo-referenciada (SIG) para relacionar espacialmente los datos estadísticos.

Para ejemplificar la importancia de contar con este tipo de información, veamos un sencillo ejemplo. Supongamos que se desea ampliar la cobertura de servicios de salud mediante el establecimiento de pequeñas unidades de salud en comunidades muy marginadas. La información estadística nos permite conocer el número de este tipo de localidades y el municipio en que se encuentran. Pero disponer de sistemas geo-referenciados abre un nuevo abanico de posibilidades para la toma de mejores decisiones. Al contar con mapas precisos de la ubicación territorial de las localidades en cuestión, se pueden establecer criterios de vecindad, densidad, concentración poblacional, o comunicaciones entre comunidades, para decidir en dónde establecer este servicio y cuántos de ellos necesitan instalarse. Este proceso conlleva mayores probabilidades de una decisión más eficaz en el uso de los recursos que otra basada en un menor volumen de información.

[4] De Janvri y Sadoulet, 2002, muestran los resultados de un estudio sobre la focalización de un programa de desarrollo de capital humano, Oportunidades, considerando el supuesto de un objetivo alternativo para tal programa social.

Para que la tarea de geo-referenciación de información se realice, es necesario un diseño que permita la obtención de ciertas variables clave durante la recolección diseño de recolección de información desarrollado por el Instituto Nacional de Estadística, Geografía e Informática (INEGI) comprende, además de estos identificadores, también los que corresponden a niveles de agregación a nivel de manzanas o cuadras (Hernández, D.; et. al., 2003).

El mapeo de información permite identificar patrones de comportamiento de los datos estadísticos que aportan información adicional al análisis de un fenómeno, pues establecen delimitaciones territoriales que relacionan el espacio geográfico con ciertos indicadores. Por ejemplo, se sabe que las entidades federativas con mayores niveles de marginación están ubicadas en el sur del país, en tanto el norte está caracterizado por menores niveles de marginación. Este tipo de comportamientos reflejan las similitudes que existen entre los individuos que conforman una población y las diferencias entre grupos poblacionales de distintas zonas.[5]

En resumen, la focalización de los programas sociales que contribuyen a la superación de la pobreza puede adoptar una serie de mecanismos de trabajo que permitan dirigir los apoyos a la población que se encuentra en situación de riesgo o vulnerabilidad. Dicha metodología comprende, entre otras cosas, la elección de las fuentes primarias de información y de los niveles de desagregación a los que se desea llevar la planeación de la estrategia, la selección de las herramientas estadísticas, matemáticas y econométricas para el procesamiento de los datos, y su vinculación espacial al territorio.

1.3 Focalización geográfica y focalización individual

La aplicación de apoyos dirigidos se puede realizar a partir de mecanismos de focalización geográfica, focalización individual (a personas u hogares) o una combinación de ambas metodologías. Adicionalmente, en ocasiones se utilizan también mecanismos complementarios que tienen como base la participación social.[6]

[5] Elbers, Lanjouw, Mistiaen, Özler & Simler 2003 exploran la contribución a la desigualdad de diversos indicadores medidos a distintos niveles de desagregación. Utilizan datos de tres países: Ecuador, Madagascar y Mozambique y combinan información de encuestas en hogares, representativas para distintas regiones y a nivel nacional, con datos de censos. En la sección 2.1 de este documento se presenta una herramienta estadística que permite valorar la importancia de indicadores socioeconómicos agregados a distintos niveles.

[6] En este documento no se desarrolla el tema de focalización comunitaria por ubicarse más allá de los aspectos relacionados con la eficacia medida desde el punto de vista estadístico y económico. Algunos autores han documentado entre las ventajas de la participación comunitaria indica que los bajos niveles de participación social que en general prevalecen, limitan los resultados de estos mecanismos.

La focalización geográfica tiene como principio fundamental la selección de zonas caracterizadas por perfiles más homogéneos en comparación con la situación que prevalece en el resto del territorio. Por ejemplo, si bien es cierto que los hogares que viven en condiciones de pobreza se encuentran distribuidos en prácticamente todo el territorio nacional, puede mostrarse que una proporción importante de ellos se ubica en las zonas más marginadas o las localidades más dispersas. En este caso, una manera de hacer llegar los apoyos a estos hogares es dirigiéndose hacia esas zonas, con el supuesto de que una más elevada proporción de quienes reciban los apoyos se encuentren en condiciones de pobreza en estos lugares, en comparación con el impacto que alcanzaría si se distribuyeran en todo el país.

En materia de política social, la delimitación de las zonas en que se focaliza geográficamente se realiza con base en la medición de la marginación o la concentración de hogares en pobreza, aunque podrían considerarse indicadores de desigualdad prevaleciente entre la población con respecto del nivel educativo, la mortalidad infantil o incluso simplemente el tamaño de las poblaciones o su etnia de origen, entre otros. De esta forma, existen programas o estrategias dirigidas de manera prioritaria o exclusiva hacia localidades o municipios que conglomeran zonas con índices de alta y muy alta marginación.

Sabemos que en las localidades rurales de alta y muy alta marginación la prevalencia de indicadores de pobreza es mucho mayor en comparación con el resto del territorio nacional. Si la totalidad de los hogares que viven en condiciones de pobreza habitaran en las localidades rurales de mayor marginación del país, la focalización geográfica de los programas sociales hacia estas zonas sería suficiente para garantizar que la población que padece condiciones de pobreza sería atendida. Sin embargo, los datos indican también que no es insignificante la proporción de población en condiciones de pobreza que habita en localidades urbanas. De la misma manera, se sabe que en las localidades rurales de mayor marginación también habitan hogares que no padecen condiciones de pobreza.

Estos hechos revelan cómo la focalización geográfica ofrece ventajas para la implementación de determinadas estrategias de política pública, pero, a su vez, subrayan la conveniencia de utilizar mecanismos de focalización individual, que incorporen información adicional cuando hacia el interior de unidades geográficas se observan condiciones de heterogeneidad, facilitando tomar en cuenta los perfiles de un cierto segmento de la población que es de interés. Es decir, para alcanzar mayor eficiencia, el mecanismo de focalización puede combinar elementos de concentración geográfica con datos más desagregados.

Un ejemplo de esta propuesta es el caso de programas para la superación de la pobreza aplicados en el medio urbano. Se sabe que en 2002 la prevalencia de pobreza alimentaria o pobreza extrema (Comité Técnico para la Medición de la Pobreza, 2002) en las zonas urbanas de México corresponde a 8.5% de los hogares. Un mecanismo de focalización geográfica dirigido simplemente hacia localidades urbanas sería poco efectivo, pues implicaría que por cada cien hogares atendidos, menos de 10 tendrían condiciones de pobreza, en tanto que los 90 restantes estarían recibiendo beneficios que no requieren de acuerdo a sus condiciones socioeconómicas. Al ser los recursos escasos esto sería en detrimento de otros hogares en condiciones de pobreza que podrían quedarse sin atención.

Un mecanismo más adecuado consiste en utilizar los datos disponibles para identificar las características de la población en unidades territoriales más pequeñas que las localidades, tales como colonias, barrios o incluso cuadras o manzanas.

Pero incluso aprovechando estas herramientas, habrá hogares en condiciones de pobreza que habiten fuera de estas unidades territoriales más pequeñas, o habrá hogares en esas zonas que no tengan tal condición y no requieran los apoyos, por lo que será necesario seguir una aplicación de técnicas de focalización individual basadas en la identificación de hogares o de personas. La definición de unidades individuales tendrá que ver con el objetivo de cada programa, con aspectos de equidad y justicia, y por supuesto con consideraciones de costos, operación y logística que no deben ser desatendidos.

Por ejemplo, en el caso de programas para el impulso a las capacidades de los individuos, se considera que el hogar es la unidad social mínima en donde se desarrollan las personas, y su desarrollo está estrechamente vinculado con el desenvolvimiento de otros miembros del hogar, por lo que la focalización individual se hace a nivel del hogar.

En la siguiente sección describiremos con más detalle cómo se aplican estas diferentes aproximaciones en tres estrategias de política pública en México.

1.4 Focalización de los programas de la Secretaría de Desarrollo Social

1.4.1 Estrategia de Micro-regiones

Para la política social del Gobierno Federal de México, es imprescindible el desarrollo local, las vías de comunicación permiten llegar a los mercados y diversos servicios públicos, como la electricidad, son necesarios para impulsar la productividad. Por ello, mediante la Estrategia de Micro-regiones se dirigen recursos para abatir los

rezagos de infraestructura que han prevalecido durante décadas en las regiones rurales más aisladas del territorio.

La idea que sustenta esta acción de política pública es que la política social debe concebirse como pieza de impulso a la competitividad. En este sentido, las personas requieren de mecanismos para adquirir activos (en un sentido amplio), pero también necesitan vivir en un entorno propicio para las actividades económicas.

La Estrategia de Micro-regiones tiene un enfoque territorial que se sustenta en la convergencia de apoyos de diversos organismos del Gobierno Federal para mejorar la accesibilidad y comunicaciones de los poblados de los municipios que conforman las micro-regiones, para fortalecer su infraestructura social básica, productiva, de educación, capacitación, abasto, salud, deporte, y para el mejoramiento de las condiciones de las viviendas. Se promueve que las inversiones sean complementadas con aportaciones de los gobiernos estatales y municipales, así como de la sociedad civil.

Las acciones de esta Estrategia se focalizan en los municipios más rezagados. De los 2 mil 443 municipios del país, 368 se clasifican como de "muy alta marginación" y 906 como de "alta marginación"; 85 de los 100 municipios menos desarrollados de México se localizan en cuatro estados: Chiapas, Guerrero, Oaxaca y Veracruz.

Para identificar las zonas de trabajo de la Estrategia, se sigue un proceso en tres pasos. En primer lugar, se incluyen todos los municipios clasificados de acuerdo al Índice de Marginación (IM) calculado por el Consejo Nacional de Población como de muy alto o alto grado de marginación.

El Índice es una medida resumen que estratifica jerárquicamente unidades territoriales, como son los municipios del país, según el impacto global de distintas carencias que enfrenta la población. Los factores de exclusión o carencia de oportunidades que se emplean corresponden al analfabetismo o la falta de educación primaria, la carencia de drenaje, servicio sanitario o agua entubada, la ausencia de energía eléctrica, la presencia de piso de tierra o un alto nivel de hacinamiento en la vivienda, así como la predominancia de bajos ingresos.

La agregación de estos factores en una variable resumen se realizó mediante el Análisis de Componentes Principales, que permite transformar un conjunto de variables o indicadores en uno nuevo y facilita una interpretación más sencilla del fenómeno original al reducir el análisis a un menor número de variables. Con este método se proyecta el espacio definido por los nueve indicadores sobre un espacio unidimensional. Para el cálculo del IM se realizaron pruebas sobre los componentes principales y se concluyó la pertinencia de sólo tomar en cuenta el primero de ellos para conformar el Índice de Marginación.

La base del Análisis de Componentes Principales es el coeficiente de correlación lineal o la covarianza que existen entre las variables utilizadas. A partir de los valores que toman estas medidas de asociación cuando se consideran todas las parejas de indicadores involucrados en el estudio, es posible definir nuevos indicadores resumen denominados componentes principales. La primera componente principal es el indicador resumen que explica la mayor heterogeneidad entre los casos; es decir, tiene la mayor varianza. La segunda componente es el indicador resumen que ocupa el segundo lugar en heterogeneidad y que no está correlacionado con el primero y así sucesivamente.

Una vez que se estimaron los coeficientes que ponderan cada una de las variables estandarizadas para obtener la primera componente principal, se obtuvo el Índice de Marginación como una combinación lineal de los indicadores estandarizados. Este Índice conlleva a una ordenación de las unidades en estudio (estados, municipios o localidades), de las más marginadas a las menos marginadas. Al ser una medición de intervalo, a la vez se pueden definir grupos de unidades de estudio de acuerdo con la marginación. Así, se establecen conjuntos homogéneos de acuerdo con el valor del Índice. La técnica empleada para llevar a cabo la estratificación fue la desarrollada por Dalenius y Hodges (1957) , y se divide el rango del Índice de Marginación en 5 subconjuntos.

El índice de marginación ha sido calculado para estados y municipios utilizando los datos censales de 1990 y 2000, así como los del Conteo de Población de 1995 (Indices de Marginación, CONAPO 1995). Es importante subrayar que el ordenamiento de estados, municipios y localidades se establece con relación al conjunto de las mismas. Esto significa que un municipio que puede tener elevados rezagos al interior de su entidad federativa, puede no ser clasificado en niveles de muy alta marginación cuando se le compara con municipios mucho más rezagados de otros estados del país.

En segundo lugar, a partir de los municipios más marginados, se constituyen micro-regiones con los municipios contiguos que además tengan similitudes étnicas, culturales y socioeconómicas. En tercer lugar, las propias comunidades y las autoridades estatales y municipales ayudan a determinar aquellas localidades que son más importantes como centros de confluencia social, productiva, comercial y de servicios (salud, educación, abasto), para conglomerados de pequeñas y dispersas localidades en las micro-regiones.[7]

La identificación de estas localidades, que se denominan Centros Estratégicos Comunitarios (CEC), permite aprovechar la influencia que determinados poblados

[7] Con base en el acuerdo por el que se modifican las Reglas de Operación del Programa para el Desarrollo Local (Microrregiones), a cargo de la Secretaría de Desarrollo Social, para el ejercicio fiscal 2004.

pueden ejercer sobre otras comunidades en un área de influencia. Algunos de los criterios de selección de localidades CEC son: tamaño de la localidad, acceso a servicios públicos e infraestructura básica, e integración con otras comunidades.

La caracterización de un contexto territorial con mayores rezagos debe permitir llevar a cabo una planeación regional más eficientes para mejorar las condiciones de vida de la población. Este tipo de focalización geográfica integra espacios, agentes sociales, mercados, y políticas públicas, en donde el territorio se convierte en el eje estructurador de las estrategias de desarrollo. Esto propicia, además, que cada Microregión participe responsablemente de su propio desarrollo y aumente las capacidades de las personas y las comunidades.

1.4.2 Programa Hábitat

Otro programa de desarrollo local es Hábitat, que apoya a la población de zonas urbanas que tienen una alta concentración de hogares en condiciones de pobreza. Su mecanismo de focalización geográfica se dirige a identificar este tipo de conglomerados de manzanas basándose en datos censales agregados a nivel de manzana.

Hábitat busca mejorar el entorno en que viven los pobres de las zonas urbanas y apoyar a hacer más competitivas a las ciudades (donde se genera 85% del PIB en México).

La competitividad territorial depende de localización y recursos naturales, pero también de otras áreas en que trabaja HÁBITAT:

- Desarrollo de infraestructura urbana y de servicios,
- Impulso para aumentar la disponibilidad de suelo,
- Despliegue de servicios de capacitación y apoyo para las y los trabajadores (Centros de Desarrollo Comunitario y Casas de Atención Infantil),
- Promoción a la cooperación entre gobiernos y sector privado (Agencias de Desarrollo),
- Fortalecimiento de la cohesión social, combatiendo la violencia, la segregación y la exclusión.

El método empleado para encontrar los conglomerados de hogares urbanos en pobreza se basa en dos ideas básicas: que es posible identificar el perfil de pobreza de los hogares mediante el análisis de diversas características que se asocian a dicha condición (factores que se conocen como “variables próximas”); y que es posible realizar este procedimiento con datos que vinculen información socioeconómica de los hogares con información sobre su ubicación geográfica, se lleva a cabo con datos del Censo de Población y Vivienda.

El proceso de focalización sigue tres fases. En la primera, se aplicó a los datos para cada hogar en el Censo del año 2000 de localidades mayores a 2,500 habitantes (y cabeceras municipales con menor número de habitantes que cuentan con la cartografía de traza urbana), un sistema de clasificación de pobreza similar al utilizado en el Programa de Desarrollo Humano Oportunidades (y que describiremos en detalle más adelante). Este sistema identifica hogares en condiciones de pobreza (en el caso particular de Hábitat, hogares en el nivel de pobreza patrimonial). Este umbral de pobreza lo constituyen aquellas personas para las cuales su ingreso resulta insuficiente para cubrir las necesidades de alimentación, salud, educación, vestido, calzado, vivienda y transporte público.

En la segunda fase, se obtuvo el número de hogares en cada manzana y cuántos de ellos se clasificaban como hogares en condición de pobreza, siguiendo los lineamientos de confidencialidad de la Ley General de Estadística e Informática. Cada manzana tiene una clave de identificación, que permite realizar análisis con diversos niveles de agregación, así como vincular su información a la cartografía de traza urbana, para definir, con herramientas gráficas, las manzanas con elevada concentración de hogares pobres.

La tercera fase consistió en conformar conglomerados de manzanas con alta concentración de hogares pobres, mediante un algoritmo de agrupación de manzanas aledañas en que aglutinaran los hogares caracterizados como pobres. El algoritmo de agrupamiento espacial siguió criterios de continuidad, densidad de hogares pobres (máxima) y radio de acción (para no tener al final conglomerados muy grandes que dificultaran un trabajo de calidad). El detalle de los procesos de estimación se describen en Hernández et. al., (2003).

El mecanismo de focalización geográfica aprovecha el hecho de que los hogares en situación de pobreza en las ciudades tienden a concentrarse en zonas y a formar conglomerados. En estos grupos de manzanas, se instrumenta un modelo de trabajo que combina el mejoramiento de la infraestructura y el equipamiento de las zonas urbano-marginadas, con la entrega de servicios sociales y acciones de desarrollo comunitario, ofreciendo especial atención a grupos en situación de desventaja, como son las jefas de familia, las personas con capacidades diferentes, los adultos mayores y las personas residentes en inmuebles o zonas de alto riesgo.[8]

[8] Hábitat, Reglas de Operación 2004, SEDESOL.

1.4.3 Programa Oportunidades

El programa Oportunidades busca el desarrollo del capital humano de la población en pobreza extrema, mediante apoyos a estas familias en zonas rurales y urbanas de becas educativas desde el tercero hasta el doceavo año de educación formal, fondos de ahorro para los jóvenes que concluyen el bachillerato, suplementos alimenticios para los menores de 5 años y las mujeres embarazadas y en lactancias, servicio médico gratuito y apoyos monetarios para mejorar las condiciones nutricionales.

Para recibir los apoyos, las familias deben cumplir con compromisos de corresponsabilidad: los niños, niñas y jóvenes deben asistir diariamente a la escuela, y las madres deben asistir mensualmente a sesiones de educación para la salud y llevar a los niños menores de cinco años a la vigilancia de su estado nutricional. La corresponsabilidad se registra cada mes y sólo con la comprobación de su cumplimiento se emiten los apoyos, que se entregan de manera directa a las madres de familia (Reglas de Operación del Programa de Desarrollo Humano Oportunidades, 2004).

Los apoyos de Oportunidades no se otorgan a petición de las personas sino después de un análisis de las condiciones socioeconómicas de las familias, tanto en el medio rural como en el urbano. En tal sentido, son elegibles principalmente las familias en situación de pobreza de capacidades, es decir, aquellas cuyo ingreso no alcanza para cubrir sus necesidades de alimentación o que, si pueden hacerlo, ya no les alcanza para cubrir sus gastos en salud y educación.[9]

El proceso para llegar a la población objetivo requiere de distintas etapas y metodologías; Oportunidades tiene dos mecanismos: uno a nivel geográfico y otro a nivel de los hogares (Orozco, Hubert, 2005).

Focalización geográfica

Se seleccionan localidades en las que operará el programa, tomando en cuenta su nivel de marginación, con base en los criterios establecidos por el Consejo Nacional de Población (CONAPO).[10] Se sigue este criterio porque en las localidades más marginadas se presentan la más alta proporción de hogares en condición de pobreza (Cruz, 1999). La ampliación del programa ha seguido criterios de atención prioritaria a las localidades más marginadas, iniciando por las del medio rural. En las localidades de

[9] Ver Comité Técnico para la Medición de la Pobeza (2002).

[10] CONAPO clasifica a las localidades en 5 estratos según su nivel de marginación.

menor marginación se identifican zonas de concentración de pobreza para dirigir las acciones del programa (Gutierrez, 2002; Parker, 2003).

Adicionalmente, para cada localidad seleccionada, se verifica el acceso y la capacidad de atención de los servicios de salud y educación básica. Esto se hace mediante algoritmos de accesibilidad dependiente de las vías de comunicación existentes y la validación de los responsables locales de estos servicios de la capacidad de dar atención al número esperado de familias beneficiarias de Oportunidades en cada localidad.

Focalización individual

Dentro de cada localidad, a su vez, se realiza un segundo paso, consistente en la identificación de las familias en condiciones de pobreza más aguda. Para ello, se realizan entrevistas a cada hogar de la comunidad para recabar datos sobre su condición socioeconómica. Estos datos son después procesados por un grupo especializado (que no interviene en la recolección de información) utilizando una metodología de puntajes basada en un Análisis Discriminante, que combina la información del ingreso del hogar con sus características socioeconómicas para formar una sola medida de la condición de pobreza del hogar.

Se tiene una clasificación inicial comparando el ingreso per cápita del hogar contra el valor del umbral de bienestar establecido, y se incorpora la información disponible sobre las características de los miembros del hogar y de su vivienda: composición y tamaño de los hogares; edad, uso de lengua indígena, escolaridad, participación laboral y ocupación de los miembros del hogar; equipamiento de las viviendas y la posesión de bienes y enseres domésticos; entre otros.

Mediante el Análisis Discriminante se busca una función matemática para clasificar a los hogares en dos grupos (en pobreza y por encima de la línea de pobreza) de acuerdo al perfil que los caracteriza (Orozco, et. al., 1999). Así, se definen, los hogares pobres y los no pobres, identificados de acuerdo con las características derivadas de los datos directamente proporcionados por los hogares (y verificados por equipos de trabajo de campo en muestras específicas).

El procedimiento que se utiliza forma parte de las herramientas de focalización individual denominadas “pruebas de medios” y “aproximación de pruebas de medios”. Estas herramientas utilizan información de carácter socioeconómico para la construcción de índices que “califican” la condición de pobreza de cada hogar y permiten, con base en esta información, determinar quiénes son susceptibles de recibir los beneficios

de un programa. Una de las principales ventajas de estas estrategias de aproximación a la condición de pobreza de los hogares consiste en que permiten valorar de manera simultánea un conjunto de indicadores que revelan las condiciones de un hogar.

En la siguiente sección de este documento revisaremos con más detalles el método descrito para la focalización y a nivel de hogares se presentán comparaciones con otras metodologías para ilustrar casos prácticos de aplicación de estas mediciones.

2. Metodología adoptada por Sedesol para la focalización de sus programas sociales

Los resultados de la medición de la pobreza elaborados por el Comité Técnico de Medición de la Pobreza (CTMP), que dan lugar a las mediciones oficiales de pobreza del Gobierno de México, proporcionan información para determinar la magnitud de la pobreza.[11] Sin embargo, su utilización en la práctica para la focalización de los programas sociales requiere elaboraciones metodológicas adicionales, pues no se cuenta con información detallada sobre el ingreso de cada miembro de la población a nivel individual que permita identificar directamente a los beneficiarios de los programas sociales; las encuestas de ingresos y gastos de los hogares captan información con representatividad nacional, pero sólo a nivel de muestras.

La información sobre el ingreso y el gasto con frecuencia es difícil de obtener, tanto por su naturaleza, como por el costo (económico y de tiempo) que representa aplicar cuestionarios especializados para captar adecuadamente estas variables. Por ello, se requieren alternativas metodológicas que aprovechen otro tipo de fuentes de datos, de más fácil recolección y más costeables.

Por ejemplo, una fuente de información nacional captada a nivel de la toda población es el XII Censo General de Población y Vivienda, levantado en el año 2000. El Censo ofrece información socioeconómica sobre todos los hogares del país, incluidos datos sobre sus ingresos. Sin embargo, el ingreso que capta esta fuente no es tan preciso como en el caso de la Encuesta Nacional de ingresos y Gastos de los Hogares (ENIGH).

Para llevar a cabo un proceso de focalización, una opción consiste en extrapolar los resultados de la medición de la pobreza obtenidos a partir de una fuente de datos muy precisa, como es la ENIGH, caracterizando a los hogares en condición de pobreza a partir de su perfil socioeconómico, en una fuente de datos a nivel de hogares.

Existen diversas herramientas estadísticas para caracterizar o clasificar a los hogares a partir de su información socioeconómica, "aproximándose" al nivel de ingresos en el que se ubican. Estos procedimientos se basan en modelar la probabilidad de que un hogar pertenezca a los grupos definidos a partir de los umbrales de ingreso que definen las líneas de pobreza.

[11] En el año 2001, la Secretaría de Desarrollo Social convocó a un grupo de expertos académicos para que de manera independiente definieran una metodología para la medición de la pobreza en México. Como resultado, en el año 2002 se publicó dicha metodología, elaborada con la información de la Encuesta Nacional de Ingresos y Gastos de los Hogares, ENIGH 2000. La metodología se basa en la comparación del ingreso per cápita del hogar con el costo de una canasta de bienes. Para el caso particular que se analiza aquí se considera el nivel de pobreza de capacidades, obtenido a partir de una canasta de bienes que considera simultáneamente las necesidades de alimentación, salud y educación de la población.

De esta forma, la información que se deriva de la mejor fuente disponible de datos para los ingresos (la ENIGH), se utiliza para ajustar modelos de probabilidad cuyos parámetros son después aplicados a otra fuente de datos. Con las probabilidades estimadas en función de las características de los hogares se puede determinar su nivel de pobreza, aún cuando no se cuenta con una medición de su ingreso tan precisa como en la ENIGH.

Este procedimiento que se ha descrito es precisamente el que se desarrolló para lograr la focalización del Programa de Desarrollo Humano Oportunidades (Orozco, et. al, 1999).[12] En el caso de este programa se utilizó un Análisis Discriminante, una herramienta estadística que resultó ser útil para su aplicación operativa (la cual se describe en la sección 1.4.3 de este documento).

En etapas más recientes de la política social, este mecanismo de focalización mediante Análisis Discriminante se aplicó en el año 2001 para la identificación de la pobreza en manzanas, áreas geo-estadísticas básicas (agebs), localidades, municipios y entidades federativas (Hernández, et. al, 2003), útiles para la focalización de varios programas sociales, como es el caso de las concentraciones de pobreza o polígonos en las zonas urbanas en donde opera el Programa Hábitat.

Existen otras alternativas para obtener mediciones de pobreza tomando como punto de referencia la medición de pobreza a través del ingreso que reporta la ENIGH. Una de ellas consiste en estimar el ingreso o el gasto de los hogares a partir de métodos de regresión lineal, utilizando como variables explicativas algunas características de la población (Skoufias, et. al, 2000). Una vez que se obtiene un modelo con ajuste adecuado, los coeficientes de la regresión se aplican a otra fuente de datos en la que se desee identificar a los hogares en condiciones de pobreza (como por ejemplo, el Censo) para generar un ingreso o gasto estimado. Posteriormente, con el ingreso estimado se determina la condición de pobreza de cada hogar comparándolo con una línea de pobreza establecida.

En los dos casos descritos, el análisis discriminante y los métodos de regresión, la esencia de la medición se basa en dos características: en primer lugar, se cuenta con información precisa sobre el ingreso, con representatividad a nivel nacional; en segundo lugar, es necesario un modelo estadístico con ajuste adecuado. Con ello es posible contar con estimaciones sobre el nivel de pobreza en otras fuentes de datos cuya capta-

[12] El primer desarrollo fue concebido para la focalización a nivel de hogar del Programa de Educación, Salud y Alimentación (PROGRESA), en combinación con los niveles de marginación a nivel de localidades. En etapas posteriores, el mecanismo de clasificación de pobreza basado en el Análisis Discriminante ha sido aplicado a otros niveles de agregación para identificar concentraciones de pobreza en las zonas urbanas.

ción del ingreso no es tan precisa, pero tienen mayor cobertura poblacional que una muestra.

El objetivo de esta sección es mostrar los resultados de la aplicación del Análisis Discriminante que actualmente se utiliza en Sedesol, en comparación con otras herramientas estadísticas para modelar la probabilidad de que los hogares se encuentren en condiciones de pobreza. Asimismo, comparar el desempeño y eficiencia de cada metodología respecto de la medición elaborada por el CTMP. Se presentan tres métodos: el Análisis Discriminante, el Modelo Logit y el Modelo Logit Multinivel. La especificación técnica y las bases teóricas de cada modelo se desarrollan en el Anexo I.

Para todas las estimaciones se utiliza la misma fuente de datos: la Encuesta Nacional de Ingresos y Gastos de los Hogares 2002 (ENIGH 2002).[13] El cuadro 2.1 muestra las variables que se utilizan para los análisis.

Cuadro 2.1
Variables utilizadas en los modelos estadísticos

En todos los modelos que se estiman en esta sección se utilizan las mismas variables, con el fin de hacer comparaciones sobre la misma base de interés. Las variables son continuas y categóricas. En el caso de estas últimas, el valor de 1 indica una condición desfavorable (*asociada positivamente*) de cada característica, y el 0 es la categoría de referencia representada por el complemento de cada variable.

1. Variable dependiente:

Pobreza de Capacidades. Es 1 si el hogar se clasifica en pobreza de capacidades de acuerdo a la medición de pobreza del CTMP; es 0 en otro caso.

2. Variables independientes:

a. Características del hogar:

i. Estrato Rural. Se toma como estrato rural a aquellos hogares que viven en localidades menores a 2500 habitantes.
ii. Piso de tierra. (0 = no; 1 =sí)
iii. Sin excusado. (0 = no; 1 =sí)
iv. Con excusado pero sin conexión de agua. (0 = no; 1 =sí)
v. Estufa de gas, lavadora, refrigerador, vehículo. (0 = no; 1 =sí)
vi. Hacinamiento. Variable continua que indica el número de miembros entre el número de cuartos del hogar.

b. Características de los miembros del hogar:

i. Sexo del jefe
ii. Dependencia demográfica. Número de miembros menores de 15 años y mayores de 65 entre el número de miembros con edades entre 15 y 65.
iii. Edad del jefe
iv. Número de niños dentro del hogar
v. Seguridad Social. Se considera un hogar con seguridad social aquel en el que al menos uno de los miembros cuenta con esta prestación.
vi. Escolaridad del jefe. Se incluyen dos variables de escolaridad: sin instrucción y con primaria incompleta.

c. Características de región:

i. Regiones. Son 14 regiones que se aplicaron a la muestra de la ENIGH, construidas con base en la experiencia operativa del Programa Oportunidades y la información registrada SIG. Las variables regionales representan la proporción (o la media) de las variables a nivel hogar.

[13] La muestra consiste de 17,617 hogares y tiene representatividad a nivel nacional, en zonas rurales y urbanas.

El Cuadro 2.2 muestra las características socioeconómicas de los hogares que se encuentran en condiciones de pobreza de capacidades en comparación con el perfil de los hogares a nivel nacional. A manera de ejemplo, mientras que el 24% de los hogares a nivel nacional se ubican en el contexto rural, el 50% de los hogares en pobreza de capacidades se encuentra en este estrato; y 30% de los hogares en pobreza tienen piso de tierra, una característica tres veces más frecuente en comparación con sólo el 10% a nivel nacional. Existe una gran diferencia en el *Ingreso neto total per cápita* promedio entre los hogares clasificados como pobres con respecto al promedio nacional, que es 5 veces mayor.

Como puede apreciarse existe una diferencia clara entre las características de los hogares dependiendo de su condición de pobreza. es precisamente esta característica de la información la que permite aproximar la medición de la pobreza mediante métodos estadísticos.

Cuadro 2.2
Características socioeconómicas de los hogares

	Hogares a nivel nacional	Hogares en pobreza de capacidades[4]
Hogares rurales[1]	23.6	50.1
Con piso de tierra[1]	9.7	30.3
Sin excusado[1]	6.7	18.1
Sin conexión de agua[1]	7.5	18.1
Sin estufa de gas[1]	13.5	39.1
Sin refrigerador[1]	23.8	58.3
Sin lavadora[1]	43.2	76.1
Sin vehículo[1]	52.8	70.4
Sin seguridadsocial[1]	59.0	86.5
Hogares con mujeres jefas de familia[1]	20.0	17.0
Hogares con jefes sin instrucción[1]	13.7	26.9
Hogares con jefes con primaria incompleta[1]	22.6	34.3
Hogares en Pobreza de Capacidades[2]	21.1	100.0
Índice de hacinamiento[2]	1.8	2.9
Edad del jefe[2]	47.1	46.7
Número de niños dentro del hogar[2]	1.0	1.8
Índice de dependencia demográfica[2]	0.7	1.1
Ingreso neto total per cápita[3]	2,328	444

[1] Porcentaje; [2] Promedios; [3] Ingreso promedio a pesos de Agosto de 2002; [4] Con base en la metodología del CTMP.

2.1 Aproximación por métodos de clasificación estadística

Los métodos estadísticos utilizan distintos supuestos y procesos iterativos para su estimación. El Análisis Discriminante, por ejemplo, basa su mecanismo de estimación

en identificar una función lineal que cumpla dos condiciones: máxima separación entre-grupo y mínima varianza intra-grupo. Intuitivamente, lo que busca es dividir a los hogares en dos grupos (un grupo de hogares en condiciones de pobreza y otro de hogares que no están en condiciones de pobreza), buscando que los hogares de un mismo grupo se parezcan mucho entre sí (mínima varianza intra-grupo) y a la vez, sean lo más distinto posibles con el grupo contrario (máxima separación entre-grupos). Las definiciones técnicas del procedimiento se incluyen en el Anexo I.

En el caso del Modelo Logit, el mecanismo de estimación se basa en el criterio de máxima verosimilitud, que intuitivamente significa identificar una solución para sus parámetros tal que la probabilidad de ocurrencia de los datos sea máxima.

Los modelos multinivel buscan incorporar información adicional para obtener mejores estimaciones del modelo logit simple, considerando que las variables que explican la probabilidad de ser pobre no son únicamente aquellas que caracterizan al hogar, sino que existen patrones de correlación entre los hogares que habitan en determinados territorios, que los hacen ser más parecidos entre sí, en comparación con hogares que habitan en otras regiones del país. De esta forma, el Modelo Logit Multinivel se estima en dos niveles: a nivel de los hogares y a nivel de las regiones.

En los tres tipos de modelos descritos la variable dependiente que se utiliza es dicotómica, toma el valor de 1 cuando el hogar es pobre de capacidades de acuerdo a su ingreso y 0 en otro caso. Los coeficientes estimados en cada caso indican la contribución que cada variable explicativa tiene sobre el resultados finales.

En el caso del Análisis Discriminante, el ajuste global del modelo se mide con una prueba F que indica si la contribución de las variables utilizadas es significativa para explicar las diferencias que existen entre los hogares clasificados en pobreza de capacidades y los que se encuentran por encima de dicha línea (establecida de acuerdo con la metodología del CTMP, en este caso, la correspondiente a pobreza de capacidades). El cuadro 2.3 muestra los resultados del Análisis Discriminante. La estadística Lamda de Wilks indica que las variables independientes explican 61% de las diferencias entre los hogares pobres y los no pobres, es decir, existen variables socioeconómicas diferentes del ingreso que pueden ayudar a tipificar a los hogares de acuerdo con su nivel socioeconómico.

Sin embargo, dada la presencia de desviaciones del supuesto de normalidad que utiliza el Análisis Discriminante (ver Anexo I), para establecer la utilidad de este modelo se utilizará otra estadística: la tasa de clasificación coincidente con la clasificación original, basada en el ingreso utilizando el método del CTMP. En este caso, la tasa es de aproximadamente 84%, esto significa que si se ignorara el ingreso de los

hogares entrevistados en la ENIGH y se quisiera identificar su condición de pobreza a partir del análisis discriminante, 4 de cada 5 hogares serían clasificados en la misma categoría de ingreso que les original (pobre de capacidades o no pobre) de acuerdo con sus características socioeconómicas. Más adelante se verá qué es lo que sucede con el hogar que sería clasificado incorrectamente, pues el método discriminante clasifica como pobres a algunos hogares cuyo ingreso está por encima de la línea de pobreza de capacidades, pero que sin embargo tienen características que se parecen mucho a las de los hogares más pobres.

Como resultado de estimar un análisis discriminante se obtiene una *función discriminante* (Cuadro 2.3). Dicha función "resume" las características del hogar, expresadas a partir de muchas variables, en una sola variable continua. Esta variable es un índice que ordena a los hogares de acuerdo a su nivel de pobreza. La función discriminante y las correlaciones de las variables explicativas utilizadas en este análisis con dicha función se muestran en el mismo cuadro. En este caso, las correlaciones mayores indican que las variables más correlacionadas con la función discriminante son: índice de hacinamiento, refrigerador, estufa de gas y piso de tierra (y así sucesivamente, de acuerdo al orden que aparece en el cuadro 2.3).

Una alternativa cuando se quiere explorar la importancia de cada variable consiste en realizar simulaciones con una variable a la vez, y con ello evaluar el resultado de la clasificación cuando se mejora alguno de los indicadores. Los resultados deben además utilizarse tomando en cuenta que existen otras variables que también se relacionan con la pobreza no incluidas en el modelo por su elevada correlación con las variables ya incluidas. En este sentido, las simulaciones que predicen mayores mejoras en la condición de pobreza son las más relevantes. Evidentemente, la mejora de cada indicador conlleva un costo distinto en términos de política pública.

Cuadro 2.3
Coeficientes de la Función Discriminante

Variable	Coeficientes de la función	Matriz de estructura*
Hacinamiento	-0.23	-0.592
Sin refrigerador	-0.48	-0.588
Sin estufa de gas	-0.66	-0.544
Piso de tierra	-0.46	-0.487
Sin Lavadora	-0.26	-0.465
Estrato rural	-0.13	-0.435
Número de niños	-0.29	-0.430
Sin seguridad social	-0.51	-0.385
Dependencia demográfica	-0.19	-0.384
Escolaridad del jefe: Primaria incompleta	-0.19	-0.353
No tiene excusado	-0.32	-0.303
Si tiene excusado pero no conexión de agua	-0.21	-0.267
Escolaridad del jefe: Sin instrucción	-0.06	-0.265
Región 8	0.48	-0.253
Vehículo motorizado	-0.13	-0.239
Región 14	1.09	0.197
Región 1	1.11	0.131
Región 16	0.87	-0.124
Región 4	1.17	0.124
Región15	0.75	-0.102
Región 11	1.00	0.075
Región 10	0.62	-0.072
Sexo del jefe	0.02	0.047
Región 12	1.03	-0.045
Región 13	1.11	0.018
Edad del jefe	-0.01	0.007
Región 7	0.92	-0.006
Región 5	0.79	-0.002
Región 6	0.80	0.001
Constante	1.010	
Lambda de Wilks	0.612	
Prueba F (29) g.l	11689415	
Correlación Canónica	0.62	
84.2% de los casos clasificados coincidentes con el método del CTMP		

* Correlaciones intra-grupo combinadas entre las variables discriminantes y las funciones discriminantes canónicas tipificadas. Variables ordenadas por el tamaño de la correlación con la función.

Al igual que en el caso de la Lambda de Wilks, la interpretación sobre la significancia de cada variable en el análisis discriminante se basa en un supuesto de normalidad y de igualdad de covarianzas intra-grupo. Dado que las variables explicativas en este caso incluyen variables categóricas, su distribución no es normal multivariada, por lo que las significancias se omiten. Sin embargo, es importante mencionar que las desviaciones al supuesto de normalidad y covarianza no afectan el poder de clasificación correcta del modelo (Hair, 1998).

Con el fin de contar con elementos de comparación del análisis discriminante, en el siguiente ejercicio se estimó un modelo logit (Cuadro 2.4) con la finalidad de buscar resultados alternativos bajo otra metodología que no requiere del supuesto de normalidad ni de igualdad de covarianzas. Los resultados indican que, excepto por la variable que indica la zona de residencia *rural-urbana* y la categoría de educación del jefe del hogar *sin instrucción*, todas las variables utilizadas son significativas, es decir su valor estadístico para explicar la probabilidad de que un hogar esté en condición de pobreza es importante. El ajuste global del modelo, medido por el estadístico LR es adecuado.

El coeficiente de determinación indica que el 40% de la variación en la información de los hogares pobres de capacidades esta explicada por las variables independientes incluidas. Este resultado es menor en comparación con el 62% de variabilidad explicada en el análisis discriminante, ya se dijo que las desviaciones de los supuestos pueden estar afectando esta estadística, por lo que la comparación de los modelos se hará en función de sus resultados de clasificación únicamente.

A partir de los coeficientes estimados para el modelo logit se obtiene la probabilidad de que un hogar se clasifique en pobreza de capacidades.[14] Adicionalmente, los coeficientes indican el sentido en el cual se altera la probabilidad cuando se cambia alguna de las características del hogar, los valores positivos indican mayor probabilidad de ser pobre en presencia de esa característica.

Cuadro 2.4
Modelo Logit

Variable	Coeficiente	Razón de Momios	Probabilidad
Rural	0.11	1.12	0.53
	(1.75)		
Con Piso de Tierra	0.35	1.42	0.59
	(4.32)**		
Sin excusado	0.34	1.40	0.58
	(3.71)**		
Sin conexión de agua	0.22	1.25	0.55
	(2.48)*		
Sin Estufa de Gas	0.63	1.88	0.65
	(8.09)**		
Sin Lavadora	0.56	1.75	0.64
	(8.99)**		
Sin Refrigerador	0.48	1.61	0.62
	(7.38)**		
Sin Vehículo	0.43	1.53	0.61
	(7.36)**		
Hacinamiento	0.34	1.41	0.58
	(15.84)**		
Sexo del jefe	-0.12	0.89	0.47
	(-1.82)		

[14] Ver el Anexo I para la definición de probabilidad.

Cuadro 2.4 (continuación)
Modelo Logit

Variable	Coeficiente	Razón de Momios	Probabilidad
Dependencia Demográfica	0.28	1.33	0.57
	(6.27)**		
Edad del jefe	0.01	1.01	0.50
	(3.77)**		
Número de Niños	0.51	1.66	0.62
	(16.37)**		
Sin Seguridad Social	1.20	3.31	0.77
	(18.27)**		
Escolaridad del jefe:			
Sin instrucción	0.07	1.07	0.52
	(0.96)		
Escolaridad del jefe:			
Con primaria incompleta	0.43	1.54	0.61
	(6.55)**		
Región 1	-1.58	0.21	0.17
	(-9.53)**		
Región 4	-1.78	0.17	0.14
	(-9.69)**		
Región 5	-0.85	0.43	0.30
	(-4.65)**		
Región 6	-0.82	0.44	0.30
	(-4.64)**		
Región 7	-1.16	0.31	0.24
	(-4.61)**		
Región 8	-0.67	0.51	0.34
	(-4.40)**		
Región 10	-0.59	0.55	0.36
	(-3.56)**		
Región 11	-1.35	0.26	0.21
	(-8.38)**		
Región 12	-1.38	0.25	0.20
	(-7.24)**		
Región 13	-1.57	0.21	0.17
	(-10.06)**		
Región 14	-1.70	0.18	0.15
	(-10.83)**		
Región 15	-0.95	0.39	0.28
	(-6.28)**		
Región 16	-1.16	0.31	0.24
	(-7.09)**		
Constante	-4.06	0.02	0.02
	(-22.00)**		
LR chi(29)	6574.75		
Log Likelihood	-5207.59		
R2 ajustada	0.39		

Casos clasificados coincidentes con el método del CTMP 86.6%
Variable dependiente: Pobreza de capacidades Estadístico t en paréntesis *Significativo al 5% **Significativo al 1%

Para mayor facilidad en la interpretación, la columna de en medio en el cuadro 2.4 muestra las razones de momios para cada coeficiente. Éstas representan el incremento relativo en la probabilidad de ser pobre respecto de un hogar con las mismas

condiciones pero que no posee esa característica. Por ejemplo, la probabilidad de estar en condición de pobreza para un hogar que reside en el estrato rural, controlando por el resto de las variables, sería 12% mayor con respecto a un hogar con las mismas características pero que habita en la zona urbana, sin embargo esta variable no resulta significativa en el caso del modelo logit. Por otra parte, la probabilidad de ser un hogar en pobreza de capacidades, dado que el hogar tiene piso de tierra es significativa y 42% mayor con respecto a un hogar que no tiene piso de tierra.

La tercera columna del cuadro 2.4 muestra la probabilidad de que un hogar promedio, que tenga la característica asociada a cada una de las variables explicativas, se clasifique como pobre. Por ejemplo, los hogares que en promedio tienen menor probabilidad de ser pobres (0.14), dadas el resto de sus características, son los que se ubican en la región 4, que corresponde a hogares de los estados de Coahuila, Nuevo León y Tamaulipas. En contraste, los hogares de la región 8, conformada por hogares de algunas regiones de Hidalgo, Puebla, San Luis Potosí o Veracruz tienen mayor probabilidad de clasificarse en condiciones de pobreza (0.34).

La capacidad de clasificación correcta del modelo logit respecto de la clasificación de pobreza del CTMP es 86.6%, en comparación con 84% del análisis discriminante. Más adelante se realizan algunos cálculos detallados para determinar la razón de esta diferencia.

Ambas metodologías identifican la contribución de cada variable a la función de probabilidad en forma no discrecional, con base únicamente en la información de los hogares y el algoritmo estadístico de convergencia de los estimadores. También establecen diferentes contribuciones entre las categorías de cada variable.

El siguiente modelo muestra una técnica para la correcta estimación de los coeficientes del Modelo Logit y su nivel de significancia cuando existen variables a nivel regional que reflejan un posible anidamiento en los datos, como en el caso anterior. El Modelo Logit Multinivel que se estima a continuación es una generalización del modelo logit, en donde se considera que además de las variables que caracterizan a los hogares existen otras variables que reflejan características del territorio en donde habitan y que influyen en su condición de pobreza, de manera que muchos hogares pueden compartir una sola característica, como estar aislados en una región marginada, o poseer infraestructura adecuada si habitan en una gran zona urbana. A estas variables se les denomina comúnmente variables de segundo nivel. A las variables del hogar se les llama variables de primer nivel y pueden variar de un hogar a otro, incluso si ambos hogares habitan en el mismo territorio. Es importante decir que si bien la ENIGH no es representativa de las 14 regiones que aquí se utilizan, pues no podríamos calcular

estadísticas descriptivas con representatividad regional, o estimar regresiones para cada región, los coeficientes de las variables que representan a las regiones resultaron significativos en el modelo logit, lo cual brinda cierta evidencia de las diferencias que existen entre distintas zonas del país.

Los resultados de la estimación para el modelo logit multinivel se presentan en el cuadro 2.5, a partir de dos modelos:[15]

i. Un modelo de un nivel, en donde las observaciones están a nivel de hogar y cuya ordenada al origen tiene un componente aleatorio.

ii. Un modelo de dos niveles, a nivel de hogar y a nivel de región, con efectos aleatorios en la ordenada al origen y en dos variables de segundo nivel (piso de tierra y estufa de gas).

En el modelo i todas las variables resultan significativas, excepto la variable que representa la categoría de educación del jefe del hogar (*sin escolaridad*), lo cual es consistente con los resultados del modelo logit presentado anteriormente. La diferencia con el modelo logit simple radica en que en este modelo se han excluido las variables indicadoras de región, en cambio, las regiones se han incluido dentro de la estructura de los datos pues se desea probar su efecto como variables de segundo nivel. La variable que indica el lugar de residencia *rural-urbano* se ha omitido considerando que no es una variable a nivel de hogar sino de localidad, en el modelo ii se incluirá como variable explicativa de segundo nivel en forma agregada.

Cuadro 2.5
Modelos multinivel

	Modelo i		***Modelo ii***	
EFECTOS FIJOS	Coef	T	Coef	T
Constante	-5.07	-35.67	-5.73	-27.57
PRURURB			2.25	3.73
Piso de Tierra	0.41	16.32	2.04	2.62
PPISO_TI			-0.90	-0.54
PBANO2			0.92	0.44
PMUJER			-8.03	-2.08
Sin excusado	0.35	13.13	0.43	15.34
Sin conexión de agua	0.26	9.86	0.25	9.23
Sin Estufa de Gas	0.69	29.38	0.11	0.96
PESTGAS			2.40	4.92
Sin Refrigerador	0.44	22.25	0.44	22.06
Sin Lavadora	0.59	31.07	0.61	31.62
Sin Vehículo	0.39	21.73	0.39	21.60
Hacinamiento	0.33	50.43	0.33	49.81

[15] En los modelos no lineales de análisis multinivel no es posible utilizar factores de expansión con el software HLM, para la estimación se expandió la base de datos a nivel poblacional y se creó a partir de ella una muestra aleatoria autoponderada.

Cuadro 2.5 (continuación)
Modelos multinivel

	Modelo i		*Modelo ii*	
EFECTOS FIJOS	Coef	T	Coef	T
Sexo del jefe	-0.11	-5.26	-0.11	-5.17
Dependencia Demográfica	0.27	19.36	0.27	18.81
Edad del jefe	0.01	10.89	0.01	10.32
Número de Niños	0.53	55.95	0.53	55.95
Sin Seguridad Social	1.19	59.98	1.19	59.92
Escolaridad del jefe: Sin Instrucción	0.02	0.90	0.01	0.22
Escolaridad del jefe: Con Primaria Incompleta	0.46	22.88	0.46	22.85
ESTIMACIÓN DEL COMPONENTE DE LA VARIANZA				
Constante	0.259	0.129		
Piso de tierra			0.119	
Estufa de gas			0.062	
COMPONENTES DE VARIANZAS Y COVARIANZAS DE NIVEL 2				
Constante-piso de tierra			0.015	
Constante-estufa de gas			-0.041	
Estufa de gas-piso de tierra			-0.010	
PROBABILIDADES				
Log (P/1-P)	-1.70	-1.82		
Probabilidad estimada	0.242	0.238		
ANÁLISIS DE VARIANZA				
Varianza within	0.41	0.41		
Varianza between	0.26	0.13		
Varianza del residual	3.29	3.29		
Varianza total	3.96	3.83		
R^2	0.10	0.11		
Valor de la función de verosimilitud	-246223.10	-243369.70		
Devianza	492446.20	486739.40		
Coeficiente de Correlación intraclase	0.073			

Variable Dependiente: pobreza de capacidades. Los resultados representan los coeficientes de la regresión logística multinivel

La forma de probar la existencia de variabilidad en los datos en un segundo nivel es a partir del Coeficiente de Correlación Intra-clase (CCI), que que se calcula para el modelo i, e indica el porcentaje de variación de los datos que corresponde al segundo nivel (en este caso las regiones).[16] El CCI indica que 7% de la variación en la probabilidad de estar en condición de pobreza de capacidades se explica por diferencias entre las regiones, aunque el porcentaje de variación es relativamente muy pequeño.

Es decir, existen patrones de correlación territorial que indican que la pobreza tiene que ver no sólo con el perfil de los hogares, sino también con las características del lugar en el que habitan. El 93% restante se explica por variables omitidas en el primer nivel, que bien pudieran estar ausentes en la información de la ENIGH.[17]

[16] Ver Bryk & Raudenbusch (2002) o Snijders & Bosker (2002).

[17] Estos resultados son consistentes con los encontrados para la medición de índices de desigualdad por Elbers, C, Lanjouw, et. al., 2003.

Dado que se han incluido las variables de primer nivel disponibles en la ENIGH, ahora es importante definir qué variables a nivel de región (segundo nivel) explican diferencias entre un territorio y otro. El hecho de que existan patrones de correlación dentro de las regiones indica por otra parte que los errores estándar del modelo logit están subestimados y por lo tanto la significancia de los coeficientes puede estar sobrestimada. Además, los valores de algunos parámetros cambian debido a la nueva especificación, aunque cabe destacar que en el modelo ii son cercanos a los estimados previamente.

La selección de variables de segundo nivel se define a partir de correlaciones de Pearson de los residuales bayesianos y de Mínimos Cuadrados Ordinarios, que indican qué variables son significativas (Cuadro 2.6). Adicionalmente, el análisis exploratorio basado en regresiones de los Residuales Bayesianos con las variables explicativas de segundo nivel (Cuadro 2.7) permiten apoyar la decisión. Los resultados de los cuadros 2.6 y 2.7 son consistentes.

Cuadro 2.6
Correlaciones Pearson de residuales y posibles variables de segundo nivel

	Bayes empíricos	*MCO*
PBANO1: Proporción de hogares sin baño por región	0.317	0.316
PREFRI: Proporción de hogares sin refrigerador por región	0.726**	0.726**
PLAV: Proporción de hogares sin lavadora por región	0.614*	0.614*
PVEHI: Proporción de hogares sin vehículo propio por región	0.369	0.369
HACINA: Índice de hacinamiento	0.678**	0.678**
DEPDEMOG: Índice de dependencia demográfica	0.682**	0.682**
EDADJ: Edad promedio del jefe	-0.082	-0.083
NINOS: Número de niños por región	0.662**	0.662**
PSS: Proporción de hogares	0.600*	0.600*
PESCJ1: Proporción de jefes de hogar sin instrucción	0.652*	0.652*
PESCJ2: Proporción de jefes de hogar con primaria incompleta	0.746**	0.746**
PRURURB: Proporción de hogares rurales	0.721**	0.721**
PPISO_TI: Proporción de hogares con piso de tierra por región	0.560*	0.560*
PBANO2: Proporción de hogares sin conexión de agua por región	0.539*	0.539*
PESTGAS: Proporción de hogares sin estufa de gas por región	0.706**	0.706**
PMUJER: Proporción de hogares con jefatura femenina por región	-0.329	-0.330

**Correlación significativa al 10% *Correlación significativa al 5%

Cuadro 2.7
Coeficientes de la regresión de residuales bayesianos sobre variables de segundo nivel

Variable	Coeficiente
Proporción de hogares con piso de tierra por región	2.87 (2.34)*
Proporción de hogares sin baño por región	2.47 (1.16)
Proporción de hogares sin conexión de agua por región	3.43 (2.22)*
Proporción de hogares sin estufa de gas por región	2.48 (3.46)*
Proporción de hogares sin refrigerador por región	2.20 (3.66)*
Proporción de hogares sin lavadora por región	1.45 (2.69)*
Proporción de hogares sin vehículo propio por región	1.29 (1.38)
Índice de hacinamiento	1.02* (3.20)*
Proporción de hogares con jefatura femenina por región	-6.01 (-.21)
Índice de dependencia demográfica	5.03 (3.23)*
Edad promedio del jefe	-0.04 (-.29)

Resultados de la regresión de residuales bayesianos sobre variables de segundo nivel para una posible inclusión en modelos subsecuentes. Estadístico t en paréntesis. *Significativo al 5%

El Cuadro 2.8 muestra las correlaciones entre las variables de segundo nivel que resultaron significativas de los análisis realizados en 2.6 y 2.7. Las variables estufa de gas y piso de tierra están muy correlacionadas con casi todas la variables restantes, como refrigerador, lavadora, índice de hacinamiento, escolaridad del jefe del hogar, etc. y, a su vez, están muy correlacionadas entre sí. Por lo tanto, dos posibles variables de segundo nivel a incluir son estufa de gas y piso de tierra.

Intuitivamente parecería claro que un hogar que tenga piso de tierra y no cuente con una estufa pueda clasificarse como un hogar que enfrenta condiciones de pobreza. Adicionalmente, el resultado del análisis derivado de los cuadros 2.6 y 2.7 indica que este tipo de hogares tienden a estar concentrados en ciertas zonas, probablemente estas variables están también reflejando otro tipo de carencias asociadas a la pobreza. Por ejemplo, podría tratarse de hogares en donde casi ningún hogar cuenta con estufa de gas porque en principio no existe siquiera la infraestructura de caminos necesaria para proveer de este combustible a la zona, de manera que la variable estufa de gas agregada a nivel regional esté reflejando también este hecho, en ausencia de una variable que nos indique la disponibilidad del servicio de distribución de gas o su precio. Como no se cuenta con esta información no es posible corroborar esta hipótesis, aunque sería deseable hacerlo. En este ejercicio se utilizarán las variables que indican la posesión de estufa de gas y la presencia de pisos de tierra como variables que permiten aproximarse a una visión regional de lo que ocurre con la pobreza. Se incluyen como proporciones en el segundo nivel.

El cuadro 2.8 muestra únicamente un análisis exploratorio de correlaciones de segundo nivel. La valoración que define finalmente las variables de segundo nivel a

utilizar es la significancia de los coeficientes de la regresión de residuales bayesianos que se muestra en el cuadro 2.9. De las variables medidas a nivel de región, sólo la proporción de hogares rurales, la proporción de jefes de familia mujeres y la proporción de hogares sin estufa de gas son significativas.

Cuadro 2.8
Correlaciones entre variables de segundo nivel

	ppiso_ti	pbano2	pestgas	Prefri	plav	Hacina	depdemog	ninos	pss	pescj1	pescj2
ppiso_ti	1.00	**0.76**	**0.83**	**0.86**	**0.89**	**0.77**	**0.61**	0.56	**0.84**	**0.82**	**0.85**
pbano2	0.76	1.00	0.73	0.75	0.71	0.68	0.47	0.44	0.56	0.62	0.63
pestgas	**0.83**	**0.73**	1.00	**0.96**	**0.89**	**0.93**	**0.66**	0.61	**0.77**	**0.86**	**0.87**
prefri	0.86	0.75	0.96	1.00	0.91	0.92	0.73	0.67	0.82	0.84	0.87
plav	0.89	0.71	0.89	0.91	1.00	0.88	0.66	0.58	0.86	0.91	0.87
hacina	0.77	0.68	0.93	0.92	0.88	1.00	0.65	0.60	0.71	0.82	0.81
depdemog	0.61	0.47	0.66	0.73	0.66	0.65	1.00	0.98	0.85	0.82	0.86
ninos	0.56	0.44	0.61	0.67	0.58	0.60	0.98	1.00	0.78	0.78	0.79
pss	0.84	0.56	0.77	0.82	0.86	0.71	0.85	0.78	1.00	0.92	0.94
pescj1	0.82	0.62	0.86	0.84	0.91	0.82	0.82	0.78	0.92	1.00	0.95
pescj2	0.85	0.63	0.87	0.87	0.87	0.81	0.86	0.79	0.94	0.95	1.00

Cuadro 2.9
Coeficientes de la regresión de residuales bayesianos sobre variables de segundo nivel

	Constante	PPISO_TI	PESTGAS
PPISO_TI	-0.93(-0.97)		0.59(0.96)
PBANO1	0.44(0.29)	-0.96(-0.78)	0.10(0.10)
PBANO2	-0.57(-0.46)		0.95(1.25)
PESTGAS	0.00(0.00)	0.13(0.23)	-0.01(-0.01)
PREFRI	-0.02(-0.03)	0.02(0.05)	0.02(0.06)
PLAV	-0.15(-0.33)	-0.28(-0.73)	0.17(0.57)
PVEHI	-0.62(-0.94)	-0.21(-0.36)	-0.20(-0.46)
HACINA	0.10(0.33)	-0.06(-0.23)	-0.03(-0.16)
PMUJER	-5.06(-1.56)		3.57(1.74)
DEPDEMOG	0.02(0.02)	-0.07(-0.06)	0.80(0.89)
EDADJ	-0.11(-1.32)	-0.12(-1.73)	
NINOS		0.25(0.40)	
PSS		-0.49(-0.81)	
PESCJ1		-0.47(-0.46)	
PESCJ2		-0.35(-0.52)	

Resultados de la regresión de residuales bayesianos sobre variables de segundo nivel para una posible inclusión en modelos subsecuentes. Estadísticos t en paréntesis.

El modelo ii, del cuadro 2.5, considera las dos variables de segundo nivel sugeridas, el resultado indica que una vez que se controla por la posesión de estufa y la presencia de piso de tierra a nivel regional, ya no resultan significativas otras variables a este nivel. Este modelo se compara con el modelo i del mismo cuadro para determinar el mejor ajuste.[18] La variaza intra-clase, es decir, dentro de cada región, es la misma en los dos modelos, pues corresponde a la varianza intra-clase de toda la muestra. Dado que se trata de un modelo logit la varianza residual puede fijarse en $n^2/3=3.29$ (Snijders, 2002).

La varianza entre-grupos, es decir, entre regiones, disminuye en el modelo ii. Esto es razonable porque este modelo incluye efectos aleatorios en las dos variables a nivel de región, que absorben parte de esta varianza. De hecho, esta reducción es de aproximadamente 50 por ciento. Como resultado, el coeficiente de determinación en el modelo ii resulta ligeramente mayor, i.e., 11% de la variación en la variable dependiente está explicada por las variables independientes, en comparación con 10% en el modelo i.

Los valores de las devianzas resultan consistentes con este resultado, pues en el modelo ii la devianza es significativamente menor, lo que sugiere mejor ajuste. El modelo ii es mejor, en él todas las variables de hogar son significativas, excepto la

[18] Encontrar una especificación adecuada tanto en primer como en segundo nivel es una tarea que se realiza probando diferentes hipótesis. En este trabajo se presentan los resultados finales de entre los numerosos modelos que se probaron.

variable *sin estufa de gas* que pierde su significancia a nivel individual al absorber su efecto el segundo nivel.

Al igual que en un modelo logit simple, en un Modelo Logit Multinivel los coeficientes no representan cambios marginales, ni probabilidades; dado que las variables de hogar interactúan con las de región, a partir de los coeficientes se puede saber si la probabilidad de ser pobre aumenta o disminuye. Para saber de qué magnitud es el cambio, es necesario realizar simulaciones.

De acuerdo a la ecuación (2.6) del Anexo I, para valores negativos de el valor de *p* disminuye y para valores positivos *p* aumenta. Así, para un hogar con todas las características favorables que vive en una zona urbana, la probabilidad de estar en pobreza de capacidades es muy pequeña. Sin embargo, si ese hogar se encuentra en una región predominantemente rural, la probabilidad de estar en condiciones de pobreza aumenta. Si el hogar tiene piso de tierra o no tiene estufa de gas, la probabilidad de encontrarse en pobreza aumenta. También hay que recordar que además de los efectos entre hogares y entre regiones, también existe un efecto aleatorio particular al hogar y a la región no explicado en el modelo, que aumenta (o disminuye) la probabilidad de hallarse en condición de pobreza.

Un rasgo más importante de los modelos multinivel es que permiten ir más allá del análisis de las variables a nivel del hogar en combinación con variables medidas en otros niveles (en este caso regiones), así como explorar el efecto que esta combinación tiene sobre la condición de pobreza. Si bien las características del hogar son importantes en el modelo, incluir variables agregadas a nivel de región demuestra que, efectivamente, existen variaciones en las condiciones de pobreza de los hogares que son explicadas por variaciones entre regiones. No considerarlas provocaría un sesgo en los estimadores y su nivel de significancia, pues en algunos casos las variaciones de segundo o tercer nivel son más importantes incluso que a nivel del hogar. En la siguiente sección se verá como afecta esto la clasificación de hogares en condición de pobreza.

En el caso de las características *piso de tierra* y *sin estufa de gas* a nivel hogar se puede ver claramente un ejemplo tanto en el modelo i como en el ii, donde la variable *piso de tierra* es significativa. Sin embargo, al incluir las variables de región: *proporción de hogares con piso de tierra en la región* (PPISO_TI), *proporción de hogares sin baño con conexión de agua en la región* (PBAÑO2) y *proporción de hogares con jefes de familia mujer en la región* (PMUJER), la variable de piso a nivel de hogar pierde parte de su significancia.

Esto indica que tener piso de tierra es un indicador de pobreza importante. Sin embargo, habitar en una región en donde muchos hogares tienen piso de tierra dismi-

nuye la importancia de la variable a nivel hogar, reflejándose como más relevante el nivel agregado pues el contexto puede provocar que incluso contar con piso de cemento o con recubrimiento no sea relevante. En este último modelo, aunque dos de las variables de segundo nivel no son significativas, la variable PMUJER sí lo es y su efecto es negativo. Esto significa que el efecto de la variable piso de tierra sobre la probabilidad de encontrase en pobreza de capacidades es menor cuando el hogar habita en una región con elevada proporción de jefas de hogar mujer, posiblemente debido a que muchas de ellas habiten en hogares con pisos de tierra.

Con respecto a estufa de gas, esta variable es significativa, pero al incluir la variable de región en el modelo ii pierde su significancia, pues es absorbida por la variable de región. Esto significa que el no tener estufa de gas en un hogar no es tan importante para indicar una situación de pobreza como el hecho de que ese hogar viva en una región donde la mayoría de las viviendas no tienen este tipo de equipamiento doméstico.

Como se puede apreciar, el Modelo Logit Multinivel aporta algo de información adicional sobre comportamientos regionales, sin embargo su estimación e interpretación son considerablemente más complejas que en el caso del Análisis Discriminante y el Modelo Logit simple. En la siguiente sección se realizan comparaciones sobre la capacidad de clasificación de cada modelo respecto de los otros, se buscará la solución metodológicamente más simple y con mejor capacidad de predicción.

2.2 Análisis comparativo de algunos métodos estadísticos: análisis discriminante, modelo logit, modelo logit multinivel

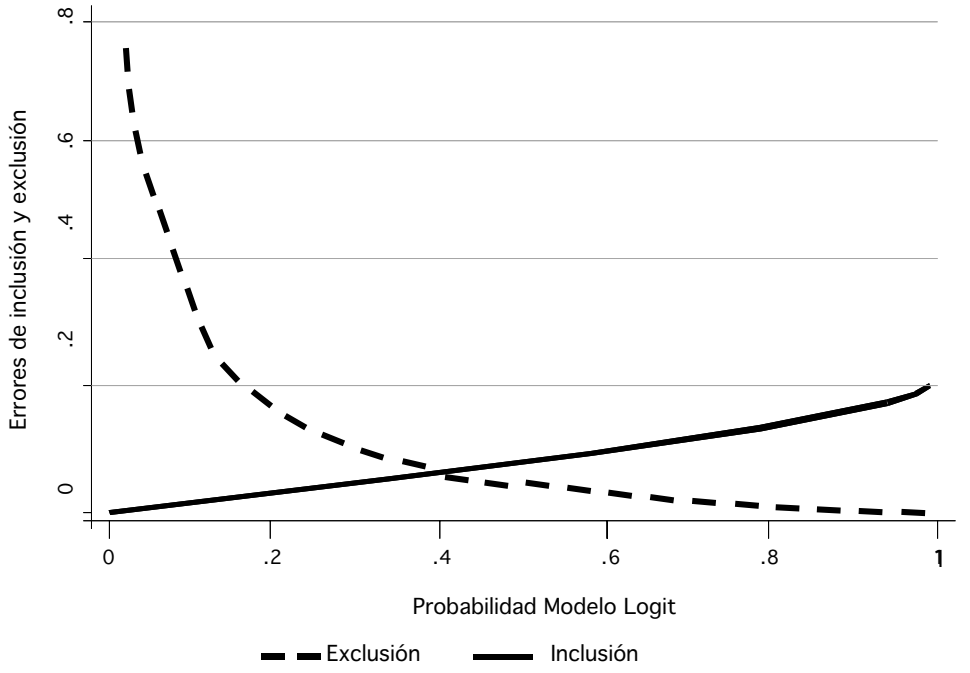
Un criterio práctico para identificar la eficiencia de las tres técnicas estadísticas utilizadas, consiste en verificar su capacidad de clasificar a los hogares de acuerdo a los datos. En esta sección se comparan los porcentajes de *clasificación correcta* de cada uno de ellos, es decir, el porcentaje de casos que son clasificados en la misma categoría (pobre o no pobre) en comparación con la clasificación del CTMP, pues en este caso dicha medición es la que consideramos el estándar para medir la pobreza.

En todos los casos los modelos predicen una probabilidad de estar en condiciones de pobreza, por default cada modelo utiliza un criterio en el cual si la probabilidad de estar en condiciones de pobreza es mayor o igual a 0.5, el hogar se considera en condiciones de pobreza. En este ejercicio, con el fin de minimizar la tasa de error[19] del

[19] La palabra error en este texto tiene el significado de error estadístico.

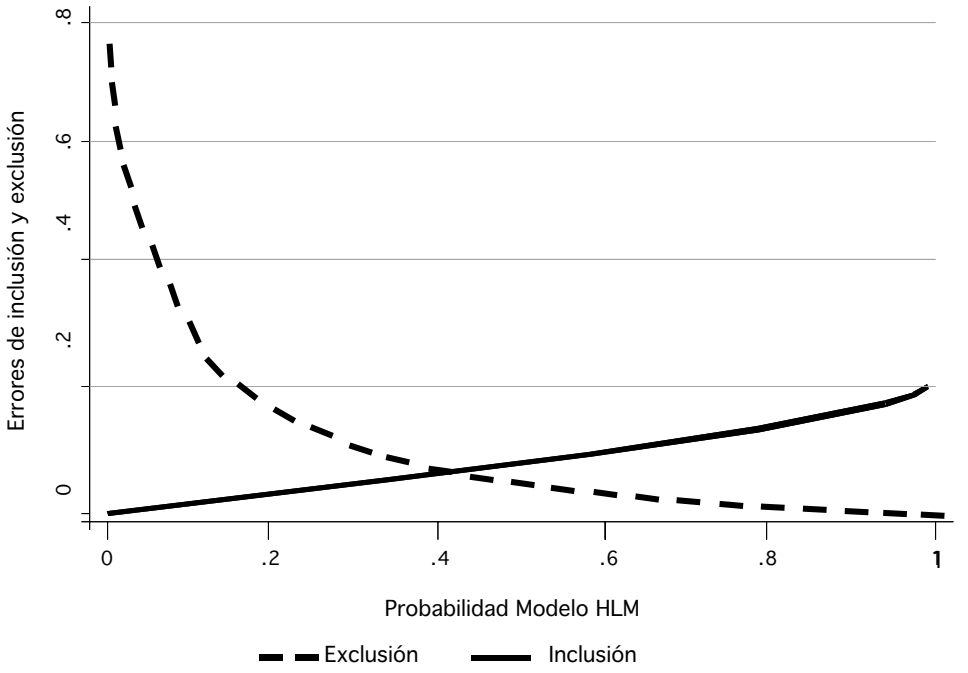
modelo logit y del modelo multinivel, se consideraron dos niveles de *corte* para las probabilidades predichas. En el primer caso se utilizó la probabilidad de 0.50 (es decir, que un hogar está en pobreza de capacidades si la probabilidad de ser pobre es mayor o igual a 0.50) y en el segundo caso, para el modelo logit se consideró una probabilidad de 0.36 (es decir, que un hogar se clasifica en pobreza de capacidades si la probabilidad que predice el modelo es mayor a 0.36). Este último criterio se tomó con base en la minimización de los errores de inclusión y exclusión que arroja el modelo para distintas probabilidades.[20] Tal y como se muestra la Gráfica 2.1, el punto donde ambos errores son mínimos es cuando la probabilidad es aproximadamente 0.36.

Gráfica 2.1
Errores de inclusión y exclusión para distintas probabilidades según el modelo Logit



[20] El error de inclusión se define como el número de hogares pobres de acuerdo al modelo Logit y que son no pobres de acuerdo a la metodología del CTMP, divididos entre el número total de hogares. El error de exclusión son los hogares clasificados como no pobres en el modelo logit y pobres de acuerdo a la metodología del CTMP, divididos entre el número total de hogares.

Gráfica 2.2
Errores de inclusión y exclusión para distintas probabilidades según el modelo multinivel



Para el modelo multinivel se utilizó el mismo criterio, pero en este caso el punto donde la diferencia entre errores es mínima es cuando la probabilidad del modelo es 0.40 (Gráfica 2.2).

Para valorar la eficacia en la clasificación de cada uno de los modelos se utilizan tres aproximaciones. En primer lugar, se realiza la comparación entre los porcentajes de *clasificación correcta* de cada modelo en comparación con el criterio del CTMP, es decir sus respectivas tasas de fuga y subcobertura. En segundo lugar, a través de índices FGT se obtienen mediciones para comparar la brecha y la profundidad de la pobreza para cada modelo. En tercer lugar, se hace un análisis empírico, presentando las características de los hogares clasificados como no pobres en relación a ciertos indicadores que intuitivamente reflejan situación de pobreza, como son el piso de tierra, la carencia de enseres, la falta de acceso a mecanismos de seguridad social, entre otros, con el fin de verificar la proporción de hogares con estas características específicas que de acuerdo a cada metodología serían clasificados como no pobres.

El cuadro 2.10 muestra la primera aproximación utilizando los resultados de la clasificación de cada modelo en comparación con la metodología del CTMP. La tasa de subcobertura derivada del Análisis Discriminante es la tasa más baja, en comparación con los otros dos modelos. Podemos afirmar que el Análisis Discriminante es el criterio más conservador, en la medida que incluye al mayor número de hogares cuyas características se asemejan a la pobreza. Con el Análisis Discriminante, la tasa de subcobertura es 27%; el Modelo Logit y el Modelo Multinivel (utilizando el criterio de minimización de las tasas de error), tienen una tasa de subcobertura de 33 por ciento. Es decir, son menos precisos para identificar a los hogares en condiciones de pobreza.

La tasa de fuga correspondiente al Análisis Discriminante es de 40 por ciento. Esto significa que 4 de cada 10 hogares clasificados como no pobres por el Análisis Discriminante, de acuerdo a sus características socioeconómicas, serían clasificados no pobres por la metodología del CTMP. Este comportamiento se presenta también en el caso del Modelo Logit y el Modelo Multinivel, especialmente cuando se utiliza el criterio para minimizar el error de clasificación que se explicó al inicio de esta sección.

Cuadro 2.10
Selección de los modelos estimados en comparación con la metodología del CTMP.

A.

Discriminante	CTMP		
	No Pobre	Pobre	Total
No Pobre		27.0%	74.6%
Pobre	40.1%		25.4%
Total	79.2%	20.8%	100.0%

B.

Logit 0.36	Comité		
	No Pobre	Pobre	Total
No Pobre		32.9%	78.7%
Pobre	34.4%		21.3%
Total	79.2%	20.8%	100.0%

C.

Logit 0.50	Comité		
	No Pobre	Pobre	Total
No Pobre		43.1%	83.9%
Pobre	26.3%		16.1%
Total	79.2%	20.8%	100.0%

D.

HLM 0.40	Comité		
	No Pobre	Pobre	Total
No Pobre		33.3%	78.8%
Pobre	34.3%		21.2%
Total	79.2%	20.8%	100.0%

E.

HLM 0.50	Comité		
	No Pobre	Pobre	Total
No Pobre		41.8%	82.8%
Pobre	29.3%		17.2%
Total	79.2%	20.8%	100.0%

En el caso de los modelos logit, las tasas de fuga son de 34% (sin utilizar los criterios de minimización del error en el logit y el logit multinivel, las tasas de fuga son menores a 30% a costo de un incremento sustantivo en la tasa de subcobertura). Las tasas de fuga entre 34% y 40% reflejan que existen hogares por encima de la línea de pobreza de ingreso, cuyas características socioeconómicas son tan similares a las de los hogares que se encuentran por debajo de la línea de pobreza, que podrían considerarse como hogares pobres.

Generalmente, cuando no se utiliza un procedimiento de minimización simultanea de las tasas de fuga y subcobertura, como el que se mostró en las gráficas 2.1 y 2.2, la minimización de la tasa de subcobertura que generan los modelos logit y logit multinivel al utilizar el criterio de predicción cuando la probabiliad es mayor de 0.5, lleva a una mayor tasa de fuga. Este es el caso de los modelos C y E del cuadro 2.10, en donde las tasas de subcobertura corresponden a 26% y 29% para el modelo logit y logit multinivel, respectivamente, a costa de un incremento de 9 y 7 puntos porcentuales en la tasa de fuga del modelo.

La utilización de distintos criterios para la predicción del modelo que se adopta tienen que ver con el riesgo que se enfrenta al incrementar el error de clasificación en alguno de los sentidos. En muchas aplicaciones se puede cuantificar el costo monetario que implica cada error, sin embargo en el caso de la focalización de un programa social influyen también otros criterios de justicia social que no son cuantificables y en algunas ocasiones están relacionados con decisiones éticas. Por ejemplo, es imposible cuantificar el costo económico que representa excluir a un hogar en condiciones de pobreza de los beneficios de un programa social, pues la falta de apoyos puede repercutir en varias esferas de la vida de las personas en el corto, mediano y largo plazo, como el desarrollo de sus capacidades, sus posibilidades de inserción al mercado laboral, la capacidad de contar con elementos para contribuir a romper el círculo intergeneracional de la pobreza, entre otros. En el caso de la focalización del programa Oportunidades y de Hábitat, las decisiones sobre la metodología utilizada, el Análisis Discriminante, fueron tomadas dando preferencia a un criterio de justicia para los hogares, con base en su perfil socioeconómico, eligiendo el modelo más conservador en términos de evitar en lo posible errores de subcobertura.

Las tasas de subcobertura y fuga sólo indican resultados con base en el número de hogares pobres y no pobres que cada uno de los modelos clasifica y el porcentaje de clasificación de ellos respecto del criterio de referencia (en este caso la metodología del CTMP), pero siempre es deseable contar con información sobre las tasas de fuga y subcobertura de la brecha y la severidad de la pobreza. Los cuadros 2.11 y 2.12

muestran estos dos indicadores considerando las medidas relativas a la profundidad y severidad de la pobreza provenientes de índices Foster-Greer-Thorbecke (FGT) que se describen en el Anexo II. En cada uno de los cuadros hay tres columnas: el índice (0) corresponde al porcentaje de clasificación en pobreza de capacidades; los índices (1) y (2) a las medidas de brecha (o profundidad) y severidad de la pobreza, respectivamente.

Tanto en la profundidad, como en la severidad, la tasa de subcobertura del Análisis Discriminante es la menor en todos los casos. En el caso de la profundidad, su tasas de subcobertura es 19.4% menor en términos relativos a la del modelo logit y mucho menor que en el caso del logit multinivel, dado que este último presenta una tasa incluso 2.3% mayor en comparación con el logit simple. En el caso de U(2), el resultado sigue siendo el mismo, pero con menores brechas relativas entre los distintos modelos.

Cuadro 2.11
Tasas de subcobertura con el esquema de ponderación FGT (cambios porcentuales con respecto al modelo Logit con probabilidad de 0.36)

	U(0)	U(1)	U(2)
Logit .36	0.329	0.094	0.080
Logit .50	0.431	0.129	0.095
	(30.77)	(37.20)	(18.92)
HLM .50	0.418	0.124	0.093
	(26.83)	(31.20)	(15.62)
HLM .40	0.333	0.097	0.081
	(1.21)	(2.31)	(1.37)
Análisis Discriminante	0.270	0.076	0.072
	(-18.11)	(-19.43)	(-9.64)

* Los números en paréntesis representan la reducción (-) o incremento (+) en las tasas de subcobertura de cada modelo, respecto al logit con punto de corte para la probabilidad igual a 0.36

En concreto, los resultados de las tasas de subcobertura del cuadro 2.11 indican que el Análisis Discriminante es el mejor modelo para identificar a los hogares más pobres entre los pobres, es decir, aquellos cuya pobreza es más profunda y más severa.

Cuadro 2.12
Tasas de fuga usando el esquema de ponderación FGT (cambios porcentuales con respecto al modelo Logit con probabilidad de 0.36)

	L(0)	L(1)	L(2)
Logit .36	0.344	0.195	0.723
Logit .50	0.263	0.149	0.781
	(-23.33)	(-23.86)	(7.99)
HLM .50	0.293	0.135	0.354
	(-14.81)	(-30.79)	(-51.02)
HLM .40	0.343	0.189	0.682
	(-0.24)	(-3.43)	(-5.69)
Discriminante	0.401	0.265	0.811
	(16.67)	(35.90)	(12.12)

El Cuadro 2.12 indica que la tasa de fuga para el modelo logit es de 34.4 por ciento. En comparación con estos resultados, el modelo logit multinivel presenta una tasa de subcobertura ligeramente menor (-0.24%), en tanto que la focalización basada en el Análisis Discriminante presenta una tasa de fuga 16.6% mayor en comparación con el logit. Esto significa que el Análisis Discriminante es el modelo más conservador, debido a que los criterios que aplica buscan evitar la exclusión de hogares con características relacionadas a la pobreza y, como se había mostrado, en este sentido tiene la menor tasa de subcobertura.

Como resultado de la comparación de los modelos, desde el punto de vista del ingreso, la identificación con base en la metodología del Comité Técnico es la mejor, ya que estos hogares pobres son los que tienen el menor ingreso, tanto total, como *per cápita* (2,300 y 444 pesos, respectivamente) y la diferencia con los ingresos de los hogares clasificados como pobres mediante los modelos estadísticos es considerable (Cuadro 2.13).

Cuadro 2.13
Ingresos por condición de pobreza según los modelos analizados

Tipo de Ingreso	Discriminante		Logit (0.50)		Logit (0.36)	
	No Pobre	Pobre	No Pobre	Pobre	No Pobre	Pobre
Ingreso corriente total	10,251	3,631	9,969	3,224	9,959	3,431
Ingreso Neto total	9,540	3,260	9,657	2,905	9,260	3,082
Ingreso corriente total per cápita	3,186	826	3,594	673	3,084	746
Ingreso Neto total per cápita	2,910	715	3,351	587	2,813	648
Tipo de Ingreso	HLM (0.50)		HLM (0.40)		Comité Técnico	
	No Pobre	Pobre	No Pobre	Pobre	No Pobre	Pobre
Ingreso corriente total	9,663	3,287	9,940	3,453	9,999	2,721
Ingreso Neto total	8,978	2,953	9,240	3,112	9,314	2,300
Ingreso corriente total per cápita	2,981	678	3,081	739	3,089	592
Ingreso Neto total per cápita	2,717	590	2,810	644	2,831	444

Aunque precisamente la intención de utilizar metodologías estadísticas tiene el objetivo de incorporar una medición más completa de la pobreza, aplicable a un concepto de focalización con buen nivel de robustez y en donde se evita al máximo la interpretación subjetiva de quien realiza el análisis, respecto de los "pesos" o "importancia" que cada variable tiene dentro de la regla de decisión. Es por eso que las tres metodologías estadísticas se desempeñan mejor cuando se valoran en conjunto las características de los hogares.

La tercera aproximación que se utiliza en este documento para determinar la eficacia de los modelos que se utilizaron consiste comparar las características socioeconómicas de los hogares identificados en condiciones de pobreza de acuerdo con cada una de las tres técnicas estadísticas (Ingreso del CTMP, Análisis Discriminante, Modelo Logit y Modelo Logit Multinivel). En este ejercicio, se considera que si los modelos permiten una adecuada focalización, el número de hogares no pobres con características socioeconómicas de desventaja (o pobreza) será lo más reducido posible. En el cuadro 2.14, se calculó el porcentaje de hogares que presentan alguna característica de pobreza, pero que sin embargo son considerados como no pobres por cada método utilizado. Por ejemplo, se observa que el Análisis Discriminante considera como no pobres al 0.8% del total de hogares con piso de tierra que existen en el país, mientras que con el método del CTMP se considera al 3.3% de estos hogares como no pobres. Por otra parte, 1.6% de los hogares sin estufa de gas, a nivel nacional, son considerados no pobres por el Análisis Discriminante; este porcentaje corresponde a 6% de los hogares cuando se utilizan los Modelos Logit y Logit Multinivel, y corresponde a 5% de acuerdo al criterio de ingreso del CTMP. En el cuadro se muestran los porcentajes para otros indicadores, como excusado, lavadora, refrigerador, vehículo, seguridad social, escolaridad, entre otros, que permiten confirmar que el menor porcentaje de "exclusión" de algún hogar cuando tiene una característica de pobreza está dado por el Análisis Discriminante.

Cuadro 2.14
Número y porcentaje* de hogares con diversas características clasificados como no pobres según técnica de clasificación

Variable	Discriminante	Logit 0.5	Logit 0.36	HLM .5	HLM .4	Comité
Piso de Tierra	0.8%	2.6%	1.7%	2.3%	1.6%	3.3%
Sin excusado o sin conexión de agua	3.1%	5.8%	4.3%	5.6%	4.4%	6.5%
Sin estufa de gas	1.6%	3.7%	2.6%	3.6%	2.8%	5.3%
Sin refrigerador	5.4%	10.6%	8.0%	10.0%	7.9%	11.5%
Sin lavadora	20.5%	27.3%	23.5%	26.4%	23.4%	27.1%
Sin vehículo	32.9%	38.9%	35.5%	37.9%	35.3%	37.9%
Mujeres jefas de familia	15.6%	17.5%	16.5%	17.2%	16.4%	16.4%
Sin Seguridad Social	34.5%	41.5%	37.2%	40.7%	37.3%	40.7%
Sin instrucción o sin primaria completa	24.4%	31.4%	27.5%	30.6%	27.5%	30.9%
I. hacinamientol. Dependencia Demográfica	1.3	1.5	1.4	1.4	1.4	1.5
	0.5	0.6	0.5	0.5	0.5	0.6
Edad del jefe	47.4	47.3	47.3	47.3	47.3	47.2
Número de niños	0.8	0.8	0.8	0.8	0.8	0.8
Niños dentro del hogar que no asisten a la escuela						
De 8 a 12 años	0.00	0.01	0.01	0.01	0.01	0.01
De 13 a 15 años	0.03	0.04	0.04	0.04	0.04	0.04
De 16 a 18 años	0.11	0.11	0.11	0.11	0.11	0.11
De 19 a 22 años	0.20	0.20	0.20	0.20	0.20	0.21

* Porcentajes con respecto al número total de hogares según la ENIGH 2002

Para ver con mayor claridad en cuánto difieren estos métodos, en el cuadro 2.15 se obtienen las razones de proporciones de hogares clasificados no pobres por cada metodología, que tienen alguna característica de pobreza, respecto de los hogares clasificados como no pobres a partir del criterio de ingreso del CTMP. Estas razones de proporciones pueden indicar el "error de medición" de los modelos, es decir, qué tan sesgados están en comparación con la pobreza medida con los procedimientos del Comité Técnico, respecto de variables importantes. De manera que si la razón de proporción es mayor que 1, la metodología tiene una margen de error mayor que la medición del CTMP, respecto de esa característica en particular. Por el contrario, si la proporción resulta menor o igual que 1, esto indica que la metodología es igual o más eficiente para detectar hogares con esa característica particular y clasificarlos como hogares pobres.

En la columna correspondiente al Análisis Discriminante, la proporción que muestra el cuadro siempre es menor que 1, lo que significa un mecanismo más eficiente. Incluso, en todos los casos, el Análisis Discriminante produce menor proporción de hogares no pobres con desventajas sociales o de escaso equipamiento, con respecto a los resultados obtenidos por cualquiera otra de las tres metodologías (pues siempre la proporción es menor para cada renglón). Por ejemplo, el margen de error del Análisis Discriminante al clasificar como hogares no pobres a hogares con piso de tierra repre-

senta sólo una cuarta parte, 0.25, con respecto a los hogares con piso de tierra que clasifica como no pobres la metodología del Comité; para el modelo Logit .50 esta proporción es 0.76 aunque puede disminuir hasta 0.52 bajo el criterio de minimización del error cuando se establece un punto de corte para la probabilidad de 0.36; aún en este caso el tamaño del error es del doble en comparación con el Análisis Discriminante. El mismo comportamiento se observa respecto del resto de las variables que se analizan.

En descargo de las diferencias que se muestran en este documento, en donde se verifica mejor aproximación de las técnicas estadísticas respecto del perfil socioeconómico, en comparación con la metodología del CTMP, debe decirse que ésta no fue diseñada como una herramienta de focalización, su principal objetivo es medir la pobreza. Hay que mencionar que en el caso del CTMP la Sedesol considera otros niveles de pobreza (pobreza patrimonial) más elevados cuya adopción podría disminuir sustantivamente las tasas de subcobertura, a costo sin embargo de incrementar también sustantivamente las tasas de fuga para la población más pobre, pero el objetivo en este trabajo se centró únicamente en la pobreza de capacidades establecida en la Sedesol.

Cuadro 2.15
Razón de proporción de hogares no pobres que presentan cada característica con respecto al porcentaje de hogares no pobres de acuerdo la medición del CTMP con esa misma característica

Variable	Discriminante	Logit .50	Logit .36	HLM .5	HLM .4
Piso de tierra en la vivienda	0.25	0.76	0.52	0.68	0.49
Sin baño o sin conexión de agua	0.47	0.89	0.66	0.86	0.68
Estufa y Refrigerador	0.42	0.85	0.63	0.81	0.64
Lavadora	0.76	1.00	0.87	0.97	0.86
Vehículo	0.87	1.02	0.94	1.00	0.93
Jefe del hogar mujer	0.95	1.07	1.00	1.04	1.00
Sin seguridad social	0.85	1.02	0.91	1.00	0.92
Jefe del hogar sin instrucción	0.79	1.02	0.89	0.99	0.89
Residencia rural	0.65	0.94	0.78	0.92	0.79
Hogares no pobres	0.91	1.03	0.96	1.01	0.97

Finalmente, como un elemento adicional para valorar el desempeño del Análisis Discriminante, la técnica de focalización que actualmente utilizan programas sociales como Oportunidades y Hábitat, se presentan las características socioeconómicas de los hogares de acuerdo a la condición de pobreza establecida por el Análisis Discriminante y por el Modelo Logit (Cuadro 2.16). Para el caso de las clasificaciones coincidentes (pobres por ambos modelos, o no pobres por ambos modelos) se observa una buena focalización de ambas técnicas estadísticas: existe un porcentaje importante de

hogares considerados en pobreza que tienen condiciones precarias y el porcentaje de hogares clasificados como no pobres con limitadas condiciones de vida es muy bajo.

Los indicadores para los casos en que las condiciones de pobreza que predice cada modelo difieren, sugieren que en términos de las características de los hogares, el modelo Discriminante identifica como pobres a hogares en los que: 21.7% de las viviendas tienen piso de tierra; 70.9% no cuentan con seguridad social; 29.9% no tienen excusado o lo tienen sin conexión de agua; 42% viven en el medio rural; 63% no tienen un refrigerador en donde conservar sus alimentos. Es decir, el Análisis Discriminante identifica hogares con perfiles de elevadas carencias, aún cuando su ingreso pueda estar por encima de la línea de pobreza de capacidades. Esta es una de las razones por las cuales se ha preferido utilizar esta técnica más conservadora para la focalización de los programas sociales.

Cuadro 2.16
Características de los hogares por condición de pobreza Discriminante / Logit .36

	Pobres AD / No Pobres L	Pobres AD / Pobres L	No Pobres AD / No Pobres L	No Pobres AD / Pobres L
Porcentaje de Hogares	4.2%	21%	74.4%	0.04%
Rural*	42.3	60.1	11.7	25.3
Piso de tierra*	21.7	37.8	1.2	0.0
Sin excusado*	13.7	23.2	1.5	13.2
Con conexión de agua*	16.2	21.8	2.8	0.0
Sin estufa de gas*	25.1	49.7	2.2	5.2
Sin refrigerador*	63.6	71.1	7.5	9.0
Sin lavadora*	73.7	86.1	28.4	60.4
Sin vehículo*	65.9	74.1	45.6	69.3
Índice de hacinamiento**	2.4	3.5	1.3	1.7
Mujeres jefas de familia*	21.3	16.7	21.7	9.0
Índice de dependencia demográfica**	0.9	1.2	0.5	0.9
Edad del jefe**	46.6	47.4	47.4	50.3
Número de niños**	1.4	2.1	0.8	1.5
Sin Seguridad Social*	70.9	92.7	47.7	100.0
Jefes sin instrucción*	23.8	32.2	7.8	7.9
Jefes con primaria incompleta*	52.0	69.1	26.1	44.8
Ingreso Corriente total per cápita**	1,074	654	3,196	1,160
Ingreso neto per cápita**	934	569	2,918	966

* Porcentaje con respecto al número de hogares por condición de pobreza; ** Promedio por hogar.

En la literatura existen múltiples referencias sobre la conveniencia de utilizar el Modelo Logit en comparación con el Análisis Discriminante cuando los supuestos de normalidad y/o identidad de matrices de covarianza no se cumplen. Sin embargo, autores como Lachenbruch (1975) o Klecka (1980) indican que el Análisis Discriminante es una técnica robusta que puede tolerar desviaciones de los supuestos.

Documentos de años más recientes demuestran la conveniencia de utilizar el Análisis Discriminante bajo la premisa de que es una técnica de clasificación, a diferencia del logit que se utiliza preferentemente cuando se quieren establecer relaciones entre variables. Muchos de esos documentos desarrollos se centran a sistemas de reconocimiento de voz y datos, así como a predicciones de riesgo financiero, como análisis de riesgos crediticios o decisiones de quiebra de empresas. Sin embargo, la utilización en el ámbito de los ciencias sociales no es tan común, por lo que hay una posibilidad de desarrollo de la técnica, cuando empíricamente resulta viable, incorporando las recientes aportaciones de textos avanzados de estadística. El caso de la focalización es un caso típico de técnicas de clasificación, particularmente en este documento se utilizan los resultados de la clasificación para determinar la viabilidad del Análisis Discriminante, así como las dificultades metodológicas que se presentan en un Logit simple. Adicionalmente se sugiere la posibilidad de explorar técnicas estadísticas con modelos multinivel con el fin de modelar con mejor precisión la significancia de las variables que se utilizan y el poder de explicación de la variabilidad de los datos.

Conclusiones

Estadísticamente es inevitable la presencia de errores en la focalización y los responsables de implementar la política social tienen que decidir entre incurrir en un error de inclusión o de exclusión cuando se elige un método. Un error de inclusión "desperdicia" recursos del programa, o hace menos eficientes sus resultados, al incluir a beneficiarios que con ingresos mayores a la Línea de Pobreza, mientras que un error de exclusión limita el cumplimiento del objetivo de reducir la pobreza pues no incorpora a los individuos que realmente lo necesitan. Cuando se compara entre métodos, aquellos con un buen desempeño en tasas de subcobertura generalmente no son tan buenos en términos de reducir la tasa de fuga.

A pesar de ello, los resultados de Coady (2002) dan evidencia de que la focalización funciona, pues ésta proporciona más recursos a los pobres que una distribución aleatoria o uniforme. Sin embargo, existe gran heterogeneidad en el desempeño de los diferentes métodos de focalización. Una de las conclusiones es que no importa qué tan bueno sea el método, lo importante para una buena focalización es la efectividad en su implementación.

La eficacia en la implementación de los métodos de focalización depende, a su vez, de las características políticas, económicas y sociales de los países donde se llevan a cabo los programas. En el caso de México, en presencia de elevada desigualdad entre la población resultan un tema indispensable a considerar.

Sin embargo, independientemente del método que se utilice, siempre es importante tomar en cuenta que la focalización debe verse como una herramienta y no como un fin. En la toma de decisiones sobre el programa a implementar deben considerarse los costos, beneficios y capacidades administrativas, teniendo siempre presente el objetivo último de la política social: mejorar y garantizar el bienestar de la población en condiciones de pobreza.

Los resultados obtenidos en este documento indican que, a pesar de que el Análisis Discriminante tiene tasas de fuga mayores, los hogares que bajo su aplicación se encuentran en condición de pobreza tienen un perfil caracterizado por limitadas condiciones socioeconómicas; además, al tener los menores niveles de subcobertura, al aplicarse en programas sociales tiene una menor probabilidad de excluir de las acciones a hogares que efectivamente requieren de los apoyos de tales programas.

Si bien existe evidencia en la literatura sobre potenciales problemas en el desempeño del Análisis Discriminante, los resultados de este documento muestran que

las desviaciones en la clasificación son en todo caso conservadoras, tendientes a evitar el error de exclusión de los hogares más pobres de los beneficios de los programas sociales. Una agenda interesante para el futuro, en la búsqueda de mejores metodologías, en todo caso deberán estar orientadas a la incorporación de variables en otros niveles de anidamiento de la información, pues los modelos multinivel muestran los cambios en la significancia de las variables utilizadas en el modelo logit, que aunque podría sugerirse como una alternativa, en todo caso lleva a resultados marginalmente distintos del Análisis Discriminante, pero no necesariamente a mejorar en forma sustantiva los resultados de la focalización.

Actualmente, existe la tentación de igualar el término de medición de pobreza con el de focalización, pero ambos objetivos son distintos y si bien guardan relación estrecha, comparar sus resultados es incorrecto, pues las metodologías de construcción de ambos tienen objetos distintos.

Hay que recordar también que la línea de pobreza utilizada para desarrollar un mecanismo de focalización representa un papel fundamental en el desempeño de cada técnica, pues deben obtenerse mediciones operativas que permitan segmentar las prioridades de atención. Mediciones de líneas de pobreza no basadas en bienes indispensables no aportan información útil para favorecer a los que menos tienen, pues incluso entre los pobres, existen hogares cuya pobreza es más intensa y profunda.

Bibliografía

Akhter U. Ahmed y Howarth E. Bouis (2002), "Weighting what's Practical: Proxy Means Tests for Targeting Food Subsidies in Egypt," Food Policy 27 (5-6): 519-40.

Amemiya, t. "Qualitative Response Models: A Survey" (1982), *Journal of Economic Literature*, 19 (4), pp.481-536.

Banco Mundial, (2004), "La pobreza en México: una evaluación de las condiciones, las tendencias y las estrategias del gobierno", Banco Mundial.

Band, Pierre, Joel Bert, John Grace y Tony Plate (1997), "A comparison between neural networks and other statistical techniques for modeling the relationship between tobacco and alcohol and cancer", *Advances in Neural Information Processing Systems*, Vol. 9, pp 967-973, MIT Press.

Bigman, David y Hippolyte Fofack (2000), "Geographical Targeting for Poverty Alleviation", *The World Bank Economic Review*, Vol. 14, No. 1: 129-145.

Buja, Andreas, Trevor Hastie y Robert Tibshirani (1994), "Penalized Discriminant Analysis", *Annals of Statistics*, Vol. 23, pp 73-102.

Buja, Andreas, Trevor Hastie y Robert Tibshirani (1993), "Flexible Discriminant Analysis by Optimal Scoring", *Journal of the American Statistical Association*, Vol. 89, No. 428, pp 1255-1270, American Statistical Association.

Castañeda, T., Lindert, K. with De la Briére, B. Fernandez, L., Hubert, C., Larrañaga, O., Orozco, M., Vizquez, R. (2005) "Designing and Implementing Household Targeting Systems: Lessons from Latin American and The United States" World Bank.

Coady, David (2000), "The Application of Social Cost-Benefit Analysis to the Evaluation of Progresa", International Food Policy Research Institute.

Coady, David y Emmanuel Skoufias (2001), "On the Targeting and Redistributive Efficiencies of Alterative Transfer Instruments", International Food Policy Research Institute.

Coady, David, Grosh, Margaret y John Hoddinott (2002), "Targeting outcomes *redux*", International Food Policy Research Institute.

Coady, D., M. Grosh y J. Hoddinott (2003), "The Targeting of Transfers in Developing Countries: Review of experiences and lessons", Mimeo. International Food Policy Research Institute.

Coady, David P. y Susan W. Parker (2004), "Combining Means-testing and Self-selection Targeting: An Analysis of Household and Program Agent Behavior", Documento de Trabajo, División de Economía, CIDE.

Comité Técnico para la Medición de la Pobreza (2002), "Medición de la Pobreza. Variantes metodológicas y estimación preliminar", *Serie: documentos de investigación I*, SEDESOL.

Conapo, (1995) "Índices de Marginación", Consejo Nacional de Población.

Conning, Jonathan y Michael Kevane (2000), "Community Based Targeting Mechanisms for Social Safety Nets," *Social Protection Discussion Paper No. 0102*, Banco Mundial.

Cortés, F., et.al (2002), "Evolución y Características de la Pobreza en México en la última década del siglo XX", *Serie: Documentos de investigación II*, SEDESOL.

Cruz, Edith Raúl Pérez y Sergio de la Vega (1999), "Geografía de la marginación y desarrollo de Progresa" en *Más oportunidades para las familias pobres. Evaluación de Resultados del Programa de Educación, Salud y Alimentación.* SEDESOL.

Dalenius, T y Hodges, J.L. (1957). "The Choice of Stratification Points", *Skandinavisk Aktuarietidskrift*, 198-203.

De Janvry, Alain y Elisabeth Sadoulet (2002), "Targeting and Calibrating Educational Grants: Focus on Poverty or on Risk?", *CUDARE Working Paper No. 985*, University of California, Berkeley.

Elbers, Chris, Peter Lanjouw, Johan Mistiaen, Berk Özler, y Kenneth Simler (2003), "Are Neighbors Equal? Estimating Local Inequality in Three Developing Countries", Banco Mundial.

Hair, J., Anderson, R., Tatham, R. and Black, W. (1998), "Multivariate Data Analysis", Macmillan Publishing Company.

Hernández, D., Orozco, M., Camacho, J. A., Vera Llamas, H., Camacho, C., Téllez, (2003), " Concentración de hogares en condición de pobreza en el medio urbano", Cuadernos de Desarrollo Humano, Secretaría de Desarrollo Social, México.

Greene, W. (2000) "Econometric Analysis", Prentice Hall 4a ed.

Gutiérrez Juan Pablo, Stefano Bertozzi y Paul Gertler (2002), "Evaluación de la identificación de familias beneficiarias en el medio urbano" en *Evaluación de Resultados de Impacto del Programa de Desarrollo Humano Oportunidades*, Instituto Nacional de Salud Pública.

Johnson, R. A., Wichern, D. W. (1998), "Applied Multivariate Statistical Analysis", Prentice Hall.

Klecka, W. R. (1980), "Discriminant Analysis", Series: Quantitative Applications in the Social Sciences, SAGE University Paper.

Maddala, G.S. (1999), "Limited-Dependent and Qualitative Variables in Econometrics", *Econometric Society Monographs No. 3*, Cambridge University Press.

Orozco, M., Gómez de León, J. y Hernández, D. (1999) "La identificación de los beneficiarios de Progresa" en *Más oportunidades para las familias pobres. Evaluación de Resultados del Programa de Educación, Salud y Alimentación.* SEDESOL.

Orozco, M., Hubert, C. (2005) "La focalización en el Programa Oportunidades de México", Unidad de la Protección Social, Red de Desarrollo Humano, El Banco Mundial, Serie de Informes sobre Redes de Protección Social.

Parker, Susan (2003), "Evaluación del impacto de oportunidades sobre la inscripción escolar: primaria, secundaria y media superior", S*erie: documentos de investigación No.6*, SEDESOL.

Raudenbush, S. y Anthony Bryk (2002), "Hierarchichal Linear Models, Applications and Data Analysis Methods", *Advanced Quantitative Techniques in the Social Sciences Series No. 1*, Sage, 2da edición.

Scott, J., (2004), ""The distribution of benefits from Public Expenditure" en: MEXICO. Public Expenditure Review. World Bank Report No 27894. Volume II.

Sharma, S., (1996), "Applied Multivariate Techniques", Wiley Publishing Company.

Skoufias E., Davis, B., Behrman, J., (2000), "Evaluación de la selección de hogares beneficiarios en el (Progresa) Programa de Educación, Salud, y Alimentación" en *Más Oportunidades para las Familias pobres*, SEDESOL.

Skoufias, E. y H. Buddelmeyer (2004), "An evaluation of the Performance of Regression Discontinuity Design on PROGRESA". *World Bank Policy Researh Working Paper 3386*, Banco Mundial.

Snijders T., Roel Bosker (2002), "Multilevel Analysis, an introduction to basic and advanced multilevel modeling", Sage Publications.

Walton, M., Lopez-Acevedo, G., (2005) "Pobreza en México, una evaluación de las condiciones, las tendencias y la estrategia del gobierno", Banco Mundial.

Anexo I. Especificación técnica de los modelos estadísticos: Análisis Discriminante, Modelo Logit, Modelo Multinivel

Las tres técnicas estadísticas que se presentan aquí permiten modelar las características de los hogares identificados en condiciones de pobreza de acuerdo a la metodología basada en el ingreso de los hogares propuesta por el CTMP. A través de estos modelos se puede estimar la probabilidad de que un hogar pertenezca al grupo de hogares en condiciones de pobreza y predecir su condición de pobreza. En este documento se realizó el ejercicio para modelar el nivel de pobreza de capacidades, pero la metodología sería la misma si se tomara cualquier otro de los niveles de alimentación o patrimonio.

De la misma manera, los tres modelos son útiles en este caso en que se desea distinguir únicamente entre dos grupos posibles: ser un hogar en condiciones de pobreza de capacidades o no serlo. Por ejemplo, si se deseara adicionalmente distinguir a los hogares en pobreza de capacidades entre los que se ubican por debajo de la línea alimentaria y los que no se tendrían tres grupos: los hogares en pobreza alimentaria, los hogares en pobreza de capacidades pero con ingreso mayor a la pobreza alimentaria y los hogares no pobres de capacidades. Las tres técnicas que se revisarán aquí permiten modelar más de dos grupos en forma simultánea si se recurre al uso de las generalizaciones de los modelos. Con el Análisis Discriminante la generalización es el análisis para tres grupos; con el modelo logit y el logit multinivel, son el Multinomial Logit.

De esta forma, en todos los casos la variable dependiente es una variable dicotómica que toma valores de acuerdo al grupo al que pertenece cada hogar (pobre de capacidades o no pobre de capacidades).

Algunas de las motivaciones para el análisis multinivel son:

a. ¿Estas características sólo se presentan a nivel hogar o también son explicativas en niveles superiores, digamos, regiones?
b. ¿Es posible que solo algunas de estas características sean explicativas a nivel regional?
c. ¿Cómo se comparan las distintas regiones en términos de estas características?

Análisis Discriminante

La mejor forma de entender esta herramienta es considerar que existen dos poblaciones $_1$ y $_2$ (población de hogares en pobreza de capacidades y población de hogares que no se encuentran en pobreza de capacidades), que se desean dividir de acuerdo con sus

sus características, representadas en un vector $x = (x_1, x_2, ...,x_k)$. El análisis discriminante permite realizar esta separación a través de la construcción de una combinación lineal entre la variable dependiente *Y* y el vector de características de la población *X*. Fisher (1936) sugiere construir una combinación lineal de la siguiente forma:

$$y = \lambda_1 x_1 + \lambda_2 x_2 + ... + \lambda_k x_k \qquad (1.1)$$

donde $\lambda = (\lambda_1, \lambda_2, ..., \lambda_k)$ son coeficientes tales que la razón de la diferencia al cuadrado entre y_1 y y_2 con la varianza de *Y* es máxima. Aquí, $\bar{y}_1$ y $\bar{y}_2$ representan el valor de la función *Y* evaluado en el valor medio del vector *X* o centroide para las poblaciones 1 y 2, respectivamente.

Si las medias de *X* en los dos grupos se representan con μ_1 y μ_2, respectivamente, y las matrices de varianzas y covarianzas $\sum_1$ y $\sum_2$. Las medias en los dos grupos en la combinación lineal de *Y* son $\lambda'\mu_1$ y $\lambda'\mu_2$. Suponiendo que $\sum_1 = \sum_2 = \sum$, la matriz de varianzas y covarianzas está dada por $\lambda'\sum\lambda$. La razón que se maximiza para cumplir el criterio sobre el que se basa el discriminante es:

$$\varphi = \frac{[\lambda'(\mu_1 - \mu_2)]^2}{\lambda'\Sigma\lambda}$$

Los coeficientes que se obtienen a partir de esta maximización están dados por:

$$\lambda = \Sigma^{-1}(\mu_1 - \mu_2) \qquad (1.2)$$

Se escogen regiones R1 y R2 tal que si la muestra cae en R1 clasificamos al individuo en *n1* y si cae en R2 lo clasificamos en *n2*. La mejor manera de definir estas regiones es utilizar un criterio que minimice el valor esperado de la proporción de observaciones clasificadas correctamente, que se denota como TMP. Sea p1 la proporción del grupo *n*1 y p2 la proporción del grupo *n*2 de la población total, la TMP se define como:

$$TMP = p_1 \int_{R_2} f_1(x)dx + p_2 \int_{R_1} f_2(x)dx$$

donde f1(x) y f2(x) son las funciones de probabilidad de las distribuciones de las características x en las dos poblaciones. Sabiendo que

$$\int_{R_2} f_1(x)dx + \int_{R_1} f_1(x)dx = 1$$

la función es minimizada si se escoge R1 y R2 tal que
p2 f2(x) < p1 f1(x) o bien

$$R1: \frac{f_1(x)}{f_2(x)} > \frac{p_2}{p_1} \qquad y\ R2: \frac{f_1(x)}{f_2(x)} > \frac{p_2}{p_1} \tag{1.3}$$

Si *x* se distribuye normal con medias μ_1 y μ_2, respectivamente y matriz de covarianza $\sum$,

$$f_1(x) = (2\pi)^{-k/2}\ |\Sigma|^{-1/2}\ \exp\left[-\frac{1}{2}\ (x-\mu_i)'\Sigma^{-1}(x-\mu_i)\right]$$

A partir de esta ecuación y tomando en cuenta la condición (3.3) se obtiene, después de tomar logaritmos, la regla de clasificación

$$(\mu_1-\mu_2)'\Sigma^{-1}x > 1n\ \frac{p_2}{p_1} + \frac{1}{2}\ (\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2)$$

Sustituyendo la definición (3.2) se obtiene

$$\lambda' x > 1n\ \frac{p_2}{p_1} + \frac{1}{2}\ \lambda'(\mu_1+\mu_2)$$

La función discriminante lineal depende de los siguientes supuestos:

a. $f_1(x)$ y $f_2(x)$ son normal multivariados.
b. Las matrices de varianzas y covarianzas son iguales (i.e. $\sum_1=\sum_2$)
c. Las probabilidades iniciales p_1 y p_2 son conocidas.
d. Las medias μ_1 y μ_2 y la matriz $\sum$ son conocidas.

Al relajar estos supuestos se obtienen diferentes casos y distintos tipos de análisis. Si se relaja el supuesto 1 la función lineal ya no es la adecuada cuando se trata de variables discretas. Al relajar el supuesto 2 la función discriminante es una función cuadrática. Si no se conocen las probabilidades iniciales (supuesto 3) se pueden utilizar las proporciones como estimadores, siempre y cuando estas provengan de una muestra aleatoria de toda la población.

Cuando no se satisface el supuesto 4 se podrían utilizar los correspondientes valores de la muestra. Sin embargo, se ha demostrado que esto produce errores, así que una alternativa es utilizar el enfoque Bayesiano. En este enfoque se supone una

distribución inicial para los parámetros y se derivan las medias posteriores de la distribución de la función discriminante verdadera, las cuales son óptimas.

Al crear una regla de clasificación se debe estimar su precisión para clasificar observaciones de futuras muestras. Esto se lleva a cabo mediante validación cruzada, la cual consiste en construir la regla de clasificación con un conjunto de datos y después utilizarla para clasificar un conjunto de datos independientes y poder estimar la tasa de clasificación correcta. Existen dos formas de llevar a cabo esta validación.

La primera consiste en dividir la información en dos submuestras generadas de manera aleatoria, una de aprendizaje y otra de validación, donde la primera se utiliza para construir la regla de clasificación y después esta regla se aplica a la submuestra de validación. La segunda involucra un proceso de dos pasos. Primero, una observación se elimina y las funciones lineales se determinan utilizando las restantes N-1 y éstas se utilizan después para clasificar a la observación eliminada. Este proceso se lleva a cabo N veces y la proporción de observaciones eliminadas que se clasificaron correctamente se utiliza como estimador de la tasa de clasificación correcta. Este método se utiliza para muestras pequeñas.

La estimación de los coeficientes de la combinación lineal brinda información adicional para la verificación del desempeño de la regla de clasificación. Algunos conceptos claves que brindan esta información son:

- Correlación canónica. Es la correlación de la función con la calificación discriminante.
- Calificación discriminante. Es el valor resultante de la aplicación de la función discriminante a los datos.
- Coeficientes discriminantes no estandarizados. Se utilizan para clasificar a partir de la función tal como se llevan a cabo predicciones en regresiones convencionales. El producto de los coeficientes no estandarizados con las observaciones da como resultado la calificación discriminante.

Una de las pruebas de significancia relevantes en este análisis son las Lamdas de Wilks. Se obtienen como resultado de una prueba ANOVA (F) de diferencia de medias. Mientras más pequeña sea la lamda de una variable independiente, esta variable contribuye más a la función discriminante. La lamda varía de 0 a 1, donde 0 significa que las medias entre grupos de esa variable son diferentes (i.e. que esa variable explica más la diferencia entre grupos) y 1 que las medias entre ambos grupos es la misma. La prueba F de las lamdas muestra si la contribución de las variables es significativa.

Modelo Logit

El modelo logit es uno de los más utilizados para el análisis de variables respuesta dicotómicas. En este tipo de modelos se considera una variable respuesta latente Y_i^*, la cual es continua y esta definida por la siguiente relación:

$$Y_i^* = \beta X_i + u_i \qquad (2.1)$$

Y_i^* no es observable. Lo que se observa es una variable dicotómica definida por:

$$Y_i = 1 \quad \text{si } Y_i^* > 0$$
$$Y_i = 0 \quad \text{e. o. c} \qquad (2.2)$$

En este caso βX_i no es $E[Y_i|Xi]$, como en un modelo de probabilidad lineal, sino $E[Y_i^*/Xi]$, en donde X es un vector de variables explicativas asociadas a la probabilidad de ocurrencia del evento (es decir, estar en condiciones de pobreza de capacidades en el caso de los resultados que se presentan en este documento). A partir de (2.1) y (2.2) se define la probabilidad de ocurrencia de la variable dependiente:

$$\text{Prob}(Y_i = 1) = \text{Prob}\,(Y_i^* > 0) = \text{Prob}(u_i > -\beta X_i)$$
$$\text{Prob}(Y_i = 1) = 1 - F(-\beta X_i) \qquad (2.3)$$

donde F es la función de distribución acumulativa de u.

Esta definición indica que los valores observados de Y provienen de un proceso binomial con probabilidades dadas por (2.3) que varían dependiendo de cada X_i . El vector de parámetros β se estima por medio de una función de verosimilitud, la cual permite encontrar el valor de la probabilidad del evento. Al maximizar esta función se obtienen valores con diferencias mínimas entre el valor estimado y el verdadero valor. Suponiendo que $u_i \sim i.i.d$ la función de verosimilitud se define como:

$$L = \prod_{Y_i=0} F(-\beta X_i) \infty \prod_{Y_i=1} [1 - F(-\beta X_i)] \qquad (2.4)$$

La forma funcional de F dependerá de los supuestos que se hagan sobre u_i en (2.1). Si la función de distribución acumulada de u_i es logística, se obtiene el modelo Logit:

$$F(-\beta X_i) = \pi_i = \frac{\exp(-\beta X_i)}{1+\exp(-\beta X_i)} = \frac{1}{1+\exp(\beta X_i)} \qquad (2.5)$$

La probabilidad para este modelo se obtiene mediante la fórmula:

$$p = \frac{\exp(\beta X_i)}{1+\exp(-\beta X_i)} \qquad (2.6)$$

En el modelo Probit se hace el supuesto de que $u_i \sim N(0,\sigma^2)$. La función de verosimilitud en este caso es:

$$F(-\beta X_i) = \int_{-\infty}^{-\beta X_i/\sigma} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{t^2}{2}\right) dt \qquad (2.7)$$

En el caso de modelo Probit se asume que $\sigma=1$. Dado que la distribución normal y la logística son muy parecidas, los resultados que se obtienen en muestras grandes utilizando (2.5) o (2.6) no deben ser muy diferentes. Ya que las varianzas de la distribución en ambos modelos son distintas, (1 en la normal y $\pi^2 / 3$ en la logística), para comparar sus resultados los coeficientes del modelo probit se multiplican por $\pi/\sqrt{3} \approx 1.8$.[21]

Otras alternativas para comparar ambos modelos son:

1. Comparar la suma de las desviaciones al cuadrado de las probabilidades
2. Comparar el porcentaje de predicciones correctas
3. Analizar los efectos marginales, i.e. las derivadas de las probabilidades con respecto a las variables explicativas dadas por:

$$\frac{\partial}{\partial X_{ik}} \Phi(X_i\beta) = \varphi(X_i\beta)\beta_k \text{ para el modelo Probit y}$$

$$\frac{\partial}{\partial X_{ik}} L(X_i\beta) = \frac{\exp(X_i\beta)}{[1+\exp(X_i\beta]^2} \beta_k \text{ para el modelo Logit.}$$

Entre las medidas más recomendadas para verificar la bondad de ajuste de los modelos Logit y Probit se encuentran las siguientes:

a. El valor de la función de verosimilitud en el máximo valor.
b. El estadístico LR. Surge de una prueba de homoscedasticidad y tiene una distribución $X^2(p)_{g.l.}$ donde p es el número de variables explicativas dentro del modelo. Si este

[21] Sin embargo, Amemiya (1981) encontró por ensayo y error que al multiplicar por 1.6 se obtienen mejores resultados. Para una explicación más detallada ver Greene [2000] y Amemiya [1981].

estadístico es mayor que el valor crítico no rechazamos H_0, y este modelo es homoscedástico.

b. El coeficiente de determinación R^2. Para el caso de variables cualitativas este coeficiente no es el mismo que se utiliza en OLS. Se han desarrollado muchos coeficientes tales como la pseudo R^2 que toma en cuenta la restricción del rango de las variables cualitativas.

Análisis Multinivel

El análisis multinivel es una metodología para el estudio de datos con patrones de variación complejos, en particular con fuentes de variación anidadas, por ejemplo estudiantes en clases o familias en manzanas o regiones. En el análisis de estos datos es importante tomar en cuenta la variación asociada con cada nivel de anidación: existe variación entre estudiantes pero también entre escuelas y uno puede llegar a conclusiones erróneas si cualquiera de esas fuentes de variación se ignora.

El análisis multinivel en el que se basa este trabajo consiste de Modelos de Efectos Mixtos, los cuales son modelos estadísticos de análisis de varianza y de regresión donde se asume que algunos coeficientes son fijos y otros aleatorios. Uno de estos modelos se conoce como Modelos Jerárquicos o HLM (Hierarchical Linear Model) y es un modelo de regresión en el cual dentro de la ecuación que define al modelo existen uno o más términos de error, según el número de niveles que existan en el análisis. En este trabajo el análisis multinivel se lleva a cabo en dos niveles. La variable dependiente es una variable de primer nivel y las variables explicativas se encuentran en ambos niveles.

Sea $i = (1,\ldots n)$ el número de individuos de primer nivel (microunidades) anidados con $j = (1,\ldots,N)$ grupos de segundo nivel (macrounidades)

La relación de primer nivel es:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij} \qquad \text{donde} \qquad r_i \sim N(0,\sigma^2) \qquad (3.1) \quad (4.1)$$

Las relaciones de segundo nivel son:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \qquad (3.2)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j}$$

El modelo en forma combinada es:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}W_j + \gamma_{11}X_iW_j + u_{0j} + u_{1j}X_{ij} + r_{ij}$$ donde se asume que:

$$E\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad Var\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} \tau_{00}\ \tau_{01} \\ \tau_{10}\ \tau_{11} \end{bmatrix} = T \qquad (3.3)$$

$$Cov(u_{0j}, r_{ij}) = Cov(u_{1j}, r_{ij}) = 0$$

- B_{0j} *y* β_{1j} son los coeficientes de nivel 1. Estos no son conocidos pero pueden ser estimados y son de 3 formas:

1. Fijos: $\beta_{1j} = \gamma_{10}$
2. No aleatorios: $\beta_{1j} = \gamma_{10} + \gamma_{11}W_j$
3. Aleatorios: $\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j}$

- $\gamma_{00,\ldots,}\ \gamma_{11}$ son coeficientes de nivel 2 y también se les llama efectos fijos.
- X_{ij} variables explicativas de nivel 1.
- W_{ij} variables explicativas de nivel 2.
- r_{ij} es el término de error de nivel 1.
- u_{0j}, u_{1j} son términos de error de nivel 2.
- σ^2 es la varianza de nivel 1
- τ_{00}, τ_{01}, τ_{11} son componentes de varianzas y covarianzas de nivel 2.

Dependiendo de la forma en los coeficientes de nivel 1 y nivel 2 se obtienen distintos submodelos, algunos de estos son:

i. Modelo ANOVA con efectos aleatorios. No hay variables explicativas en nivel 1 ni en nivel 2. Se le llama modelo incondicional.

$$Y_{ij} = \gamma_{0j} + u_j + r_{ij} \qquad \text{donde} \qquad \beta_{0j} = \gamma_{0j} \qquad (4.4)$$

ii. Modelo ANOVA con intercepto aleatorio

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + u_j + r_{ij} \qquad (4.5)$$

iii. Modelo con coeficientes aleatorios. Todos los coeficientes varían aleatoriamente.

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}W_j + \gamma_{11}X_iW_j + u_{0j} + u_{1j}X_{ij} + r_{ij} \qquad (4.6)$$

Al igual que en los métodos de regresión convencionales, cuando se trata de variables dependientes discretas no es recomendable utilizar regresiones lineales debido a:

1. Ya que el rango de la variable dependiente es restringido una regresión lineal puede arrojar resultados fuera de este rango.
2. Para las variables discretas existe generalmente una relación natural entre la media y la varianza de la distribución, es decir, para una variable dicotómica *Y* con una probabilidad *p* cuando toma valores de 1 y *1-p* para valores iguales a 0 la media es:

   $E[Y] = p$

   Y la varianza es

   $Var(Y) = p(1-p)$

La varianza está determinada en la media, lo que lleva, en el análisis multinivel, a una relación entre los parámetros con efectos fijos y los parámetros con efectos aleatorios. El modelo de regresión logística toma en cuenta estos elementos. En este modelo el resultado del individuo *i* en el grupo *j* es expresado como la suma de la probabilidad en este grupo más un residual individual.

$Y_{ij} = P_j + R_{ij}$ ó específicamente

$$Logit(Pj) = 1n\left(\frac{p}{1-p}\right) = \gamma_0 + U_{0j} + R_{ij} \qquad (4.7)$$

En este caso también se cumple $E(r_{ij}) = 0$ pero $Var(r_{ij}) = (n_i\,(1\text{-}n_i))^2\,\tau_0^2$ donde n_i se define como en la ecuación (2.5) y $\tau_0^2 = \text{var}(u_{0j})$

La ecuación (4.8) es equivalente a la (4.5) pero en esta ecuación se asume que la varianza de nivel 1 es constante. En (4.8) los grupos tienen diferentes varianzas. Por lo tanto, el parámetro σ^2 debe ser interpretado como la varianza promedio de todos los grupos.

El modelo logístico también puede formularse como un modelo threshold en el cual la variable dependiente *Y* es el resultado de una variable continua no observada $\breve{Y}$.[22] El modelo threshold especifica que:

$$Y = Y - \begin{cases} 1 \text{ si } \breve{Y} > 0 \\ 0 \text{ si } \breve{Y} \leq 0 \end{cases}$$

Pruebas de especificación y medidas de bondad de ajuste

- Para probar la hipótesis de que un parámetro sea igual a cero: H_0: $\gamma_h = 0$ utilizamos el estadístico t.

[22] Ver sección 1.

• Como ya se mencionó en la sección anterior existen diferentes definiciones del coeficiente de determinación (R^2). Una de estas definiciones esta basada en una representación threshold y se asume que las variables explicativas son variables aleatorias.

Sea un modelo ANOVA con intercepto aleatorio

$$\breve{Y}_{ij} = \gamma_0 + \sum_{h=1}^{h} \gamma_{hij} X_{hij} + U_{oj} + R_{ij}$$

La parte fija del modelo es

$\breve{Y}_{ij} = \gamma_0 + \sum_{h=1}^{h} \gamma_{hij} X_{hij}$ con varianza igual a σ_F^2. Asimismo, la varianza del intercepto es *var (Uoj)* = $\tau^2{}_0$ y la varianza de residual de nivel uno es *var(Rij)*= $\sigma_R^2 = n^2/3$

La varianza total de $\breve{Y}_{ij}$ es var($\breve{Y}_{ij}$) = $\sigma_F^2 + \tau_0^2 + \sigma_R^2$. La parte explicada es σ_F^2 y la no explicada $\tau_0^2 + \sigma_R^2$. De la no explicada τ_0^2 es de nivel dos y σ_R^2 de nivel uno. Entonces la proporción de la varianza explicada es:

$$R^2 = \frac{\sigma_F^2}{\sigma_F^2 + \tau_0^2 + \sigma_R^2}$$

• El grado de semejanza entre micro unidades que pertenecen a un macro unidad se expresa por medio del *coeficiente de correlación entre clases* (ICC por sus siglas en inglés).[23] Existen diferentes definiciones de este coeficiente, dependiendo de los supuestos en el diseño muestral. Para un modelo logístico puede ser definido de dos maneras:

$$1.\ \rho_1 = \frac{\text{varianza poblacional entre macro unidades}}{\text{varianza total}} = \frac{\tau^2}{\tau^2 + \sigma^2}$$

$$2.\ \rho_1 = \frac{\tau^2}{\tau_0^2 + \pi^2 / 3}$$

Se interpreta como la proporción de la varianza que es explicada por el grupo. Es un coeficiente de correlación porque es igual a la correlación entre valores de dos micro unidades escogidas a aleatoriamente dentro de una macro unidad escogida también aleatoriamente. Estas dos definiciones son diferentes y producen resultados diferentes.

[23] Traducción de *intraclass correlation coefficient.*

• Devianza. Cuando se estima por medio de Máxima Verosimilitud los resultados de la estimación reportan también el valor de verosimilitud. A partir de éste se puede obtener la Devianza, que es *-2 ln(valor de verosimilitud)*. Este estadístico puede considerarse una medida de falta de bondad sin embargo en muchos modelos no se pueden interpretar sus valores directamente. Se comparan diferencias en valores de la devianza para varios modelos con el mismo conjunto de datos, i.e. se puede construir una hipótesis H_0: $D_0 = D_1$, en la cual no existe una diferencia significativa en las devianzas. El estadístico correspondiente se define como:

$d = D_0 - D_1 \sim X^2_{(m1 - m0)g.l.}$ donde m_0 es el número de parámetros del modelo 0 y m_1 el número de parámetros del modelo 1.

• El *ICC* puede ser 0 o positivo. A partir de este estadístico se puede llevar a cabo una prueba de no diferencia entre grupos o ρ_I. Esta es una prueba ANOVA F cuyo estadístico es:

$$F = \frac{\tilde{n} S^2_{between}}{S^2_{within}} \sim F(N-1, M-N)g.l. \qquad \text{donde,}$$

$$S^2_{within} = \frac{1}{M-N} \sum_{j=1}^{N} \sum_{1=1}^{nj} (Y_{ij} - \overline{Y}_j)^2 \qquad S^2_{between} = \frac{1}{\tilde{n}\,(N-1)} \sum_{j=1}^{N} n_j (Y_{.j} - Y\,..)^2$$

$$\tilde{n} = \frac{1}{N-1} \left\{ M - \frac{\sum_j n_j^2}{M} \right\}$$

n_j: número de individuos (micro unidades)
N: número de macro unidades

$$M = \sum_j n_j$$

La especificación del modelo es una de las partes más difíciles de la inferencia estadística. Hay dos consideraciones importantes que deben ser maximizadas conjuntamente: estadística y sustantiva. El propósito de la especificación es llegar a un modelo que describa los datos observados de manera satisfactoria desde el punto de vista estadístico e interesante para el fenómeno de estudio sin complicaciones innecesarias. Es decir, si se observan resultados consistentes en donde las metodologías complejas contribuyen sólo marginalmente al logro del objetivo inicial, se preferirá un modelo más sencillo en pro de la claridad y transparencia de la técnica utilizada.

Anexo II. Tasas de Subcobertura y Fuga

Para saber la precisión de los modelos con respecto a la profundidad y a la severidad de la pobreza se utilizó un esquema basado en las medidas de pobreza FGT. La tasa de subcobertura se calculó por medio de la siguiente fórmula:

$$U(\alpha) = \left(\frac{1}{N_{PC}} \right) \sum_{i=1}^{q} \left(\frac{z - i_i}{z} \right)^{\alpha}, \qquad (1.1)$$

Donde N_{PC} es el total de hogares clasificados como pobres de acuerdo al criterio de ingreso; q es el número de hogares clasificados como no pobres según el modelo de comparación y como pobres según el ingreso; z es la línea de pobreza, i_i es el ingreso corriente per cápita del i-ésimo hogar y α es el peso de la severidad de la pobreza. Cuando $\alpha = 0$, esta expresión representa la tasa de subcobertura (igual a la calculada en el cuadro IV.5), si $\alpha = 1$ se le asigna mayor peso a los que se encuentran más alejados de la línea de pobreza y cuando $\alpha = 2$ se le asigna mayor peso a los hogares que se encuentran más alejados de la línea de pobreza y que además tienen un ingreso más bajo.

Siguiendo la misma lógica, la tasa de fuga se define como:

$$L(\alpha) = \left(\frac{1}{N_{ME}} \right) \sum_{i=1}^{q} \left(\frac{i_i - z}{z} \right)^{\alpha}, \qquad (1.2)$$

Donde N_{ME} son los hogares identificados como pobres por el correspondiente modelo estimado, q es el número total de hogares no pobres según el criterio de ingreso y pobres según el modelo a comparar.

Anexo III. Regiones definidas para la estimación de los modelos estadísticos

Región 1	Baja California Baja California Sur Coahuila Chihuahua Durango Sonora
Región 4	Coahuila Nuevo León Tamaulipas
Región 5	Sinaloa Sonora
Región 6	Aguascalientes Guanajuato Jalisco San Luis Potosí Zacatecas
Región 7	Durango Jalisco Nayarit Sinaloa
Región 8	Guanjauato Hidalgo Puebla Queretaro San Luis Potosí Tlaxcala Veracruz
Región 10	Campeche Quintana Roo Tabasco Yucatán

Región 11	Guanajuato Jalisco Michoacán Querétaro
Región 12	Colima Guerrero Jalisco Michoacán Oaxaca
Región 13	Guanajuato Guerrero México Michoacán Morelos Querétaro
Región 14	Distrito Federal México
Región 15	México Morelos Oaxaca Puebla Veracruz
Región 16	Chiapas Guerrero Morelos Oaxaca
Región 18	Chiapas

“La focalización como estrategia de política
pública”, de Daniel Hernández F.,
Mónica Orozco C. y Sirenia Vázquez B.,
serie: *Documentos de Investigación*, 25
se terminó de imprimir en noviembre de 2005.

El tiraje consta de 2,000 ejemplares.

Contigo
es posible

SECRETARÍA DE
DESARROLLO
SOCIAL

SEDESOL