Cuantificando a la clase media en México en la primera década del siglo XXI: un ejercicio exploratorio

Rodrigo Negrete Prieto
y Miriam Romo Anaya

El propósito del presente ensayo es mostrar que aún sin tener una definición apriorística, estructurada, cerrada y consumada de lo que es *clase media* o sin esperar a un consenso al respecto, es posible identificar qué tramo y magnitud del espectro social en México le podría corresponder. Lo anterior, siempre y cuando se tomen en cuenta ciertos avances metodológicos que muestren brechas entre grupos de hogares, resultado de un proceso estadístico de estratificación, ello en lugar de que el investigador preestablezca exógenamente umbrales a partir de cualquier criterio ajeno a los patrones de similitud y diferencia entre esos hogares en la base de datos de La Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH). Esto es posible hacerlo, además, con información disponible a la que usualmente no se le presta la misma atención que al ingreso corriente. El ejercicio se realiza con datos del final de la primera década del presente siglo y contribuye a dar respuesta a la pregunta: ¿es México un país de clase media?

Palabras clave: clase media en México, clases sociales, pobreza, métodos multivariados, conglomeración, estratificación, filosofía bayesiana, primera década del siglo XXI, Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) 2010.

The purpose of this article is to show that even without an a priori, structured, closed, and final definition of middle class, or without having a consensus about it, it is possible to identify what part and to what extent of the social spectrum in Mexico, the middle class could correspond to. To do so are considered certain methodological advances able to show gaps resulting from a households' stratification statistical process, instead of pre-establishing thresholds which ignore the similitude and difference patterns among groups of households in the data base of the National Income-Expenditure Survey. This can be done with available information neglected in comparison with data such as current income which normally draws most of the attention on regard these kind of studies. The article has been written with data focused at the end of the first decade of this century and contributes to answer the question: Is Mexico a middle-class country?

Key words: middle class in Mexico, social classes, poverty, multivariate statistical methods, conglomeration, stratification, Bayesian philosophy, Twenty century first decade, National Income-Expenditure Survey 2010.

Recibido: 4 de abril de 2014
Aceptado: 25 de septiembre de 2014

Nota: los comentarios y puntos de vista de los autores no necesariamente reflejan los de la institución en la que laboran; los autores agradecen a Lilia Guadalupe Luna Ramírez y Benito Durán Romo su colaboración tanto en el proceso de investigación como en la integración del artículo.

AURORA, CO November 16, 2005-/Hyoung Chang/Getty Images

1. Introducción

En los últimos 15 años tanto analistas internacionales como nacionales han subrayado la importancia de que en países en desarrollo clasificables como economías emergentes consoliden una clase media. Algunos han puesto el énfasis en que la expansión de la misma no sólo es un reflejo del crecimiento económico, sino que, a su vez, lo impulsa, dando lugar así a un círculo virtuoso; por ejemplo, se señala que la clase media tiene una mayor capacidad para diferir la gratificación inmediata, por lo que sus integrantes entienden —o intuyen— la importancia del ahorro y de la propiedad, así como la de invertir en acumulación de capital humano; a lo anterior se suma su papel como consumidores que encabezan una demanda interna de mejores productos, generados, distribuidos y promovidos con mayor eficiencia e imaginación, propiciando con ello también la inversión y nuevos nichos de mercado (Banerjee y Duflo, 2007). Otros van más allá del imperativo económico, y valoran a la clase media como un factor de estabilidad y de consolidación de las democracias por tender a ser más proclive a aceptar un *ethos* meritocrático y liberal[1] creando incentivos para que las diversas ofertas políticas se corran más hacia el centro y se le apueste menos a la estridencia demagógica o a los extremismos.[2]

Particularmente, en México se ha señalado que la clase o las clases medias son un factor esencial

1 Parece que bien podría ser debatido por los historiadores de las turbulencias políticas de las décadas de los 20 y los 30 del siglo XX en Europa y, en particular, de los estudiosos del surgimiento y consolidación del fascismo. No deja de ser sintomático que la ecuación clase media = liberalismo la promuevan economistas de habla inglesa cuyas naciones no pasaron por esa experiencia.

2 Habrá que ver, a su vez, si esa vocación centrista que se le atribuye a las clases medias no llegará a un punto en que sea refutado por el ascenso del movimiento *Tea Party* en los Estados Unidos de América, donde la polarización y la estridencia en el discurso político ha sido evidente en los últimos años o, también, el ascenso de los partidos de extrema derecha al otro lado del Atlántico como consecuencia de la crisis del 2008 y sus secuelas en el marco de las tensiones que introduce, en las soberanías nacionales, la pertenencia a la Unión Europea.

en la transición hacia la democracia no sólo porque tienden a ser menos pasivas y mejor informadas que otros segmentos de la sociedad, sino porque obligan al sistema político a ubicarse en la lógica y los imperativos de los derechos universales en vez de quedar estancado en la lógica de derechos negociados, pactados y administrados por el Estado con clientelas o corporaciones, lo que marca el cambio de *pueblo* a *ciudadanía* en el discurso y, en general, en el imaginario político de los actores sociales e institucionales (Zuckerman, 2010).

Es evidente que esa tensión entre la vieja y la nueva lógica se mantiene aún viva, dejando su impronta en los conflictos en la vida política del México contemporáneo, de modo que no sólo la prédica desde los organismos internacionales, sino también esa ansiedad en torno a una posible transición inconclusa ha propiciado un vivo interés sobre si el país ya es uno mayoritariamente de clase media (lo que en el contexto equivale a establecer si las condiciones objetivas para llevar a buen puerto la larga transición mexicana están dadas o no). De la Calle y Rubio (2010) tuvieron el enorme mérito de tomar el toro por los cuernos y abrir el debate en México con una toma de posición sugestiva y audaz al respecto.

Los autores de este ensayo creemos que hay aún un largo camino por andar para dar una respuesta definitiva o categórica. En todo caso, nuestra conclusión es que resulta prematuro proclamar al país como uno de clase media y que una cautela al respecto ha de mantenerse: que no es lo mismo ilustrar una serie de cambios que se han incorporado en la vida de casi todos los segmentos sociales con el hecho de que el sello, distintivo de ese cambio, sea que uno de esos grupos se ha vuelto mayoritario. No nos interesa vincular por lo demás esta discusión a aquella otra de que las clases medias sean una condición *sine qua non* para consolidar una democracia que merezca ese nombre, con todo y sus atributos sistémicos de estabilidad y capacidad autocorrectiva.

El ensayo está centrado en un terreno más básico, que es explorar una veta metodológica y de información que contribuye a identificar la magnitud posible de ese grupo. Nos percatamos, a su vez, que todavía queda mucho por delante para decir al respecto, dado que si bien al hablar de clase media el tópico se ubica en un terreno sociológico, a la hora de tratar de saldar el asunto de una manera práctica la balanza se inclina a criterios que tienden a satisfacer más a los economistas. Este sesgo es inevitable, y en este ensayo no lo superaremos del todo dado que históricamente la estadística oficial en México se ha hecho con un mayor énfasis en lo económico que en lo social, pese a los enormes avances al respecto de los últimos 10 años en la incorporación de distintas temáticas (género y violencia, notablemente).

Una clase social supone tanto condiciones objetivas como subjetivas: nivel de vida junto con códigos y valores compartidos (López-Calva, Rigolini y Torche, 2011); sin embargo, se necesitarían desarrollar encuestas que manejaran ambos planos simultáneamente y a profundidad, y lo que hay en México y el resto del mundo es que la exploración económica y sociológica tienden a diseñar y adoptar sus propios instrumentos de captación. Creemos, también, que no es saludable el desdén mutuo entre el enfoque económico y el sociológico, y que ambas perspectivas deben poner atención en los pros y contras de las variables que privilegian en sus respectivos análisis: el pragmatismo del economista al abordar el tema le viene bien al sociólogo tanto como la imaginación del segundo al primero.

Revisando la literatura sobre el tema, lo que llama la atención es la tendencia entre los economistas a centrarse en un aspecto, factor o en una variable. Esto se puede ver, por ejemplo, en el enfoque marxista de propiedad de los medios de producción que identificaba a la clase media con el *petit burgeois* o pequeño propietario y que, de acuerdo con su drama conceptual, estaba destinado a desaparecer o volverse irrelevante para despejar la arena del conflicto entre los dos segmentos irreductibles y polares: la burguesía y el proletariado. Si bien hoy en día serán pocos los que suscriben tal análisis, mismo que no previó el surgimiento

de una clase media asalariada y masificada (a la par que la *terciarización* de las economías y la presencia de más instituciones y mediaciones sociales entre los agentes de la producción), los economistas actuales —y en especial aquéllos dentro de los organismos internacionales— siguen mirando hacia una sola variable —ahora el ingreso— como el dato que lo determina todo. Reducir el asunto a un solo principio —sea conceptual o estadístico— no deja de ser una tentación.

Max Weber, en la medida en que pudo atestiguar las vertientes de desarrollo por donde fluía la sociedad moderna, introdujo un abordaje que se hizo cargo de su complejidad: la perspectiva sociológica propiamente dicha y, con ella, un enfoque que cabría denominar multivariado, donde entraron en juego tres dimensiones: a) nivel económico, b) poder/jerarquía y c) prestigio/estatus. Es así que el nivel educativo, la calificación ocupacional y los valores, códigos o expectativas son ahora parte de la paleta de colores y matices. En la Sociología contemporánea (Bourdieu, 2001) se define una multidimensionalidad de las clases sociales en términos de capitales: económico, cultural/informacional, social o recursos basados en conexiones y pertenencia grupal y, no menos importante, capital simbólico. Es justo decir que esta objeción a la monocromía economicista no sólo ha sido enfáticamente señalada en la actualidad por parte de sociólogos (Goldthorpe y McKnight, 2006), sino también por economistas como Atkinson y Brandolini (2011); en particular, estos últimos han apuntado el problema de que las divergencias en cuanto a las magnitudes, umbrales o cotas de referencia existentes entre los diversos análisis en la literatura sólo centradas en el ingreso pueden llevar a conclusiones encontradas sobre si la clase media crece o se contrae en un periodo determinado.

El pragmatismo podría estar del lado de los economistas si se considera lo inasible que pueden ser nociones que en un marco teórico suenan bien, como tener valores o códigos compartidos, pero que de manera empírica son difíciles de identificar, mientras que la consistencia conceptual sería el consuelo sociológico; sin embargo, ambos enfoques tienen problemas en aquello que pareciera ser su fuerte: la variable ingreso es comprensiblemente la más subdeclarada en las encuestas en hogares en México y en el mundo, y nada garantiza que esa subdeclaración sea consistentemente proporcional entre los segmentos sociales. Es bien sabida la brecha que hay entre la información declarativa que captan las encuestas de ingresos por una parte y las estimaciones del ingreso disponible que se establecen por los métodos de la contabilidad nacional por la otra, así como lo difícil o arbitrario que puede resultar ajustar la distribución de la variable que reportan las fuentes primarias haciendo uso de la magnitud absoluta de esa brecha (Leyva-Parra, 2005).

La estratificación meramente por ingresos también puede llevarnos a más estratos de los que sociológicamente tiene sentido: pensemos en un hogar en el que los hijos en edad adulta inician su vida aparte bajo otro techo. Su nivel de ingresos puede ser menor que el del hogar donde crecieron, pues están en la primera fase de su inserción laboral, ¿sólo por ello el primer hogar y el segundo pertenecen a clases sociales distintas? Pero el sociólogo enfrenta asimismo sus problemas cuando la dinámica demográfica o los ciclos de vida están presentes y es posible que se le revierta de igual forma el problema: el primer hogar y el segundo pueden diferir no sólo en ingresos, sino en estatus ocupacional y educacional, ¿pertenecen entonces a la misma clase social o no?, ¿comparten los mismos valores, expectativas y referentes simbólicos?, ¿hacen uso de los mismos mecanismos de sociabilidad? Tal vez sí, tal vez no, y en muchos casos hay diferencias de grados más que rupturas categóricas. También, al enfrentar el problema concreto de la identificación de clases medias en áreas rurales, uno no siempre encuentra, por ejemplo, la complejidad jerárquica ocupacional que se da por hecho en el medio urbano: la perspectiva compleja no siempre se satisface en todos los ámbitos o éstos no complacen al total de sus criterios.

Es inevitable, entonces, emprender la medición empírica de la clase media con nociones intuitivas que se desprenden de la información disponible

más que con conceptos estructurados de manera plena y, en este ensayo, ciertamente no presumiremos tener definiciones acabadas, aunque sí el haber hallado una estrategia estadística prudente para dar una respuesta; pero, además, tampoco creemos que necesariamente haya una ventaja particular en tener conceptos plenamente estructurados al abordar este tema. Es el momento de formular una visión deflacionaria a ese respecto, como lo es también el revalorar la intuición como vehículo cognitivo de la existencia social a la que pertenecemos.

En el desarrollo de este ensayo estaremos más cerca de un enfoque bayesiano que platónico, pues el problema de los conceptos prístinos que pertenecen al mundo de las ideas es, justamente, esa precesión frente a la realidad: no están dispuestos a recibir sorpresas de ella o a ser recalibrados; la realidad imperfecta está ahí sólo para ilustrarlos, de modo que la relación entre lo conceptual y lo empírico se vuelve una carretera de una sola vía. Por otra parte, al navegar en el universo social, la intuición resulta irrenunciable, como llegó a señalar el mismo Bourdieu. Si bien en las ciencias naturales la intuición no siempre resulta buena consejera —no es intuitivo que la Tierra gire alrededor del Sol—, en el universo social, en cambio, no podemos salir a decir que una noción borrosa como la de clase media es una ilusión que merece ser disipada o que su verdadero significado se sitúa más allá de la experiencia del ciudadano común (lo que Bourdieu llama la tentación epistemocéntrica del investigador). Por impreciso que sea el término clase media, hace sentido justamente por formar parte de un sistema compartido de representación del paisaje social: nos habla de una forma de vida que todos entienden a qué se refiere, aún sin ser definida de manera plena. Ya lo decía Karl Popper: "la realidad está hecha tanto de relojes precisos como de nubes difusas y es inútil preguntar por qué es así". Wittgenstein reflexionaría, por su parte en sus *Philosophische Untersuchungen* (*Investigaciones filosóficas*), sobre esa naturaleza dual, precisa e imprecisa, al mismo tiempo, del lenguaje humano, y que lo habilita como herramienta flexible para interactuar con nuestro mundo recibiendo su retroalimentación.

En este ensayo abordaremos a la clase media, antes que nada, como un nivel de vida a partir de un conjunto de variables de gasto cuya presencia apunta a una existencia social más allá de la subsistencia, que lo mismo reflejan expectativas, una inversión de cara al futuro o acceso al crédito que disfrute del entorno, y veremos a dónde ello nos lleva y qué tanto conecta con otras variables que denotan instrucción, estatus o jerarquía. Dejaremos que las variables cuantitativas de arranque hablen por sí mismas sin predeterminar exógenamente las magnitudes o umbrales a los que deban llegar, de modo que éstos no serán una premisa, sino un resultado de la metodología seguida y que es en lo que radica, en buena medida, el aporte de esta propuesta. Configuraremos, entonces, el universo constituido por el conjunto de hogares del país —nuestras unidades primarias de observación y análisis— y en él señalaremos (para utilizar una vez más la terminología de Bourdieu) un "espacio de probabilidades de afinidad" donde se puede articular o estructurar eso que todos llamamos clase media.[3] Dicho espacio queda cuantificado en un número de hogares, y la magnitud que acumula, así como algunas de las características identificadas en tal conjunto, serán descritas.

Así, en lo que sigue, la segunda sección subraya en qué se distingue la metodología de cuantificación aquí utilizada de otras que, hoy por hoy, son la corriente dominante en la literatura sobre el tema; la tercera abunda en el procedimiento seguido; la cuarta comenta los resultados y los contrastes que más llaman la atención entre los subconjuntos de hogares clasificados respectivamente como de clase baja, media y alta; el quinto, y último apartado, es para reflexiones finales.

3 Es por eso que, para nosotros, descripciones de clase media como la que a continuación se cita nos parecen tan válidas como cualquier otro intento de definición sofisticada: "The middle class is more than income bracket (…) has come to mean having a secure job; a safe and stable home; access to health care; retirement security; time off vacation, illness (…) opportunities to save for the future; and the ability to provide a good education for one's children..." (U.S. Department of Commerce. *Middle Class in America*, January 2010).

2. Tipos de mediciones y la problemática de los criterios exógenos

En el texto introductorio se ha señalado la predilección de la mayoría de los economistas —y, en particular, de aquéllos en los organismos internacionales— por la variable ingreso. Más adelante se retomará una crítica a esta preferencia de la que estos profesionales, más que nadie, debieran estar al tanto considerando, sobre todo, que la variable que utilizan para cuantificar la clase media no es el ingreso en general, renta de los hogares o riqueza sino, específicamente, el ingreso corriente, de suyo sujeto a fluctuaciones en la vida de un mismo hogar. En esta sección se pasará primero revista a los tipos de metodología dominantes y la problemática que guardan en común lo que permite entender por qué, en este ensayo, se optó por otro camino metodológico, así como por el tipo de variables que se eligieron como punto de partida.

En general, la cuantificación estadística de la clase media tiene dos vertientes, la que utiliza criterios relativos y la que emplea los absolutos:

- Mediciones relativas: rangos determinados por rangos estadísticos. Éstas se desdoblan, a su vez, en aquellas que eligen un rango de los percentiles de la distribución de hogares de acuerdo con su percepción de ingreso corriente; por ejemplo, el conjunto de hogares comprendidos entre los percentiles 20 y 80 (Easterly, 2001) y las que establecen como referencia la mediana de la distribución del ingreso, estableciendo un rango alrededor de la misma —por ejemplo entre 75 y 125% del valor de dicha mediana— de modo que aquellos hogares cuyo ingreso corriente se sitúe entre estos valores se determina que pertenecen a la clase media (Birdsall, Graham y Pettinato, 2000; Pressman, 2011).
- Mediciones absolutas: rangos determinados por valores monetarios. Ésta es la otra gran vertiente, la cual, en principio, tal pareciera se presta mejor para hacer comparaciones internacionales y por la que optan, sobre todo, economistas vinculados al Banco Mundial (BM) dada su necesidad de amplias panorámicas. En las mediciones absolutas, por lo general, se determina un referente monetario fijo que lo mismo puede proporcionarlo la línea de pobreza del BM —1.25 dólares del año 2005, ajustados por el poder de paridad de compra (PPC) que corresponda a cada país— o, alternativamente, se adoptan las líneas de pobreza o las medianas —sea de ingresos o de gastos— de dos países contrastantes con el objetivo de tener un valor monetario mínimo y otro máximo, definiendo así un rango fijo que aplicará, en adelante, a todo el conjunto de naciones bajo estudio (con el correspondiente ajuste por PPC en cada caso). Los hogares en el rango serán aquellos que califiquen como clase media, país por país. Es así que, bajo esta vertiente, se pueden obtener rangos absolutos que van de 2 a 10 dólares diarios (Banerjee y Dulfo, 2007), de 2 a 13 dólares (Ravallion, 2008), de 10 a 20 dólares (Bussolo, De Hoyos y Medvedev, 2009) o de 10 a 100 dólares (Kharas y Gertz, 2010). Las variaciones en los últimos tres rangos dependen de la línea de pobreza del país referente que se elija (Estados Unidos de América, Italia o Luxemburgo como cotas superiores; un *pool* de naciones en desarrollo, Brasil o Portugal, respectivamente, como cotas inferiores).

  En la vertiente de medición absoluta destaca el estudio de López-Calva y Ortiz-Juárez (2011), quienes adoptan un enfoque de vulnerabilidad a la pobreza menor a 10% de probabilidad (0.1 de 1) de incurrir en ella para así establecer qué monto exacto de ingreso corriente corresponde a ese nivel de riesgos y obtener la cota absoluta mínima (10 dólares PPC); la cota absoluta máxima de 50 dólares se establece como el ingreso que promedian los percentiles 95 de los tres países de estudio (Chile, México y Perú). A primera vista, podría pensarse que éste es un método híbrido porque no utiliza un referente de distribución para establecer la cota inferior y, en cambio, sí uno para determinar la cota superior, pero

finalmente es un método absoluto porque el rango 10-50 dólares así obtenido lo aplica por igual en los tres casos (con el correspondiente ajuste por PPC).

La diversidad de los rangos arriba referidos no es casual. Al repasar este rápido inventario, podemos estar seguros de que mientras más simple sea el criterio para establecer un rango (por ejemplo ±25% del valor de la mediana del ingreso corriente) más arbitrario éste será: pero el común denominador (con la excepción de cómo se fija el valor inferior del rango en el estudio de López-Calva y Ortiz-Juárez) es una altísima dependencia —*ex ante*— del resultado con la decisión inicial tomada. Mientras más cercana quede una conclusión con respecto a sus premisas, o más las refleje, más tautológica será y, por ende, menos información proporciona (Popper, 1962)

- Un método alternativo: no preestablecer rangos. Una aproximación distinta sería abordar lo que ha de separar a los hogares de clase media de los demás, no como un punto de partida sino como un resultado al que se llega. Es por ello que en esta exploración se adopta una metodología de conglomeración —de cuyas características más adelante se abundará— porque la resultante nos muestra un espectro de brechas en el universo de hogares que no había necesidad de fijar de antemano y que nos guía de ahí en adelante en el análisis en vez de obviarlo.

Podría pensarse que esta estrategia es una nueva modalidad dentro de la vertiente de medición relativa porque el subconjunto de hogares que en el espectro se alejan de los otros no tendrían necesariamente las mismas características en México que, digamos, los que así se aíslen en Estados Unidos de América o en Europa Occidental, si fuera el caso. Pero el punto es justamente ése: qué se separa de qué en cada contexto. Por el contrario, al aplicar un rango fijo o preestablecido, ese rango pudiera señalar cotas al interior de un país que, por sí mismas, no corresponden a las diferencias subyacentes en su espectro social careciendo así de significado: simplemente un sello o molde que desde fuera se imprime en el paisaje y que la huella que deja no nos dice —por ser determinada de manera exógena— si lo que demarca es o no una diferencia arbitraria o simplemente sobrepuesta (no olvidemos que cuando se habla de clases sociales se expresa, ante todo, de algo que resalta en su contexto). Si entonces la diferencia no brota del paisaje mismo, estaremos hablando de personas que tienen tantos ingresos o gastos preestablecidos, pero no de mucho más que eso: no habremos avanzado, seguiremos en realidad reproduciendo nuestro punto de partida desde el principio hasta el final.

2.1 ¿Ingresos o gastos?

En cuanto a la variable de punto de partida, es importante señalar que no todos los estudios mencionados han elegido los ingresos como referencia. Banerjee y Dulfo (2007), así como Kharas y Gertz (2010), decidieron enfocarse en el gasto como un mejor *proxy* del consumo de los hogares, y ésa es una elección que nosotros también aquí adoptamos. Cuando se trabaja con el ingreso corriente de las encuestas, hay que estar al tanto de su naturaleza fluctuante, pues no sólo se refiere a los ingresos del trabajo —que de suyo varían cuando el perceptor tiene un trabajo independiente o también en aquellas modalidades ocupacionales que combinan una percepción fija y otras, como comisiones o bonos de productividad—, sino que, asimismo, se integra con ingresos recurrentes que provienen de la renta de la propiedad física o financiera (los intereses de una cuenta bancaria) y transferencias (por ejemplo, remesas) que no necesariamente suponen flujos de magnitud constante. El riesgo, entonces, es que si nos quedamos con el ingreso corriente y umbrales o cotas predeterminadas para decidir si se pertenece o no a tal o cual clase social, las familias con ingresos corrientes cercanos a los valores máximos y mínimos pueden estar entrando y saliendo de una condición de clase todo el tiempo.

El consumo no se ajusta con la misma intensidad y velocidad; no por nada, en la literatura económi-

ca se han formulado tesis, como la del ciclo de vida (Modigliani,1963), la cual parte del hecho de que el consumo siempre será más estable que el ingreso y el ahorro a lo largo de la vida de los individuos y las familias o, también, la hipótesis del ingreso permanente (Friedman, 1956), donde su mejor estimación no la proporcionan los ingresos corrientes actuales, sino una combinación de éstos con los ingresos esperados, reflejándose en realidad más cercanamente dicho ingreso permanente en los niveles de gasto y consumo.

En otras palabras, el ingreso corriente no es la variable más indicada para identificar un nivel de vida porque es más incierto inferir desde ahí ciertas invariantes que caracterizan al mismo. Si pensamos que el término *clase* tiene sentido porque resulta una condición relativamente estable en el paisaje social, sin duda es el gasto la variable en la que hay que enfocar la atención porque no depende sólo de un flujo sino, asimismo, de un *stock*, que es la riqueza de los hogares, variable esta última que se escapa de las fuentes de información de las que, por lo pronto, normalmente se disponen para hacer estos ejercicios de cuantificación de la clase media.

2.2 Variables de punto de partida utilizadas en este ejercicio exploratorio

Se objetará, sin embargo, que la variable gasto también puede quedar sujeta a grandes fluctuaciones —por ejemplo, puede incluir gastos funerarios, así como de la salud y hospitalarios— y es aquí, entonces, donde debemos enfocarnos no en el gasto en general, sino en rubros de gasto no contingentes pero, sobre todo, que tengan significado intuitivo.

Así, es importante que estas variables nos hablen de algo más que la subsistencia; qué tan intensamente se participa de vínculos sociales o del mundo exterior; qué relevancia se le da a presentarse frente a él; y qué tanto se invierte en ocio e información, en darle mantenimiento a las posesiones, así como en adquirir activos y en gastos derivados de la calidad de los mismos, variables todas que queden expresadas en valores per cápita. Lo anterior también puede ser complementado por información cuantitativa útil centrada en algunos activos físicos o sus características que resultan inocultables y que las encuestas de ingresos y gastos captan sin dificultad. En suma, la idea es trabajar con un *mix* de variables que combinan información de gasto corriente con algunos rubros que apuntan a un vínculo con *stocks* o riqueza (activos).

Haciendo uso de la información que en este sentido proporciona como ninguna otra fuente la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH), en su levantamiento del 2010,[4] fueron seleccionadas las variables que se muestran en la tabla 1 como punto de arranque de la exploración.

Si bien se pudieron haber incluido otras variables de gasto, se considera que éstas se vinculan de manera más clara, dentro de la multiplicidad de variables que ofrece la ENIGH 2010, a los tipos de capital ya mencionados que esbozara Bourdieu: económico, informacional, social o fincado en conexiones (gastos en regalos otorgados) e incluso simbólico y de estatus (como los gastos en cuidados personales). Una ventaja no menor es que algunas de ellas se relacionan con gastos fijos del hogar de modo que son menos dependientes de su demografía y composición. El número de 17 variables no fue preestablecido: simplemente es la cantidad que se considera relevante tanto bajo una perspectiva económica como sociológica.

No sólo la selección específica de variables de gasto marca diferencia con los otros estudios ya mencionados que sí hacen uso del dato, sino también un enfoque multivariado para hacerlo; ello porque, de entrada, el combinar información de activos físicos y gastos nos pone ante métricas dife-

4 Las ENIGH en México —que ha levantado el Instituto Nacional de Estadística y Geografía (INEGI)—, al igual que sus equivalentes en el resto del mundo, son las encuestas de hogares más utilizadas en este tipo de estudios, y es en su información en la que enteramente se basa este ensayo. La edición 2010 de la ENIGH comprendió una muestra de 30 169 viviendas, lográndose entrevistas en 28 513, en las cuales habitaban 28 968 familias; esto garantizó a la muestra representatividad nacional, con desglose urbano y rural.

Tabla 1

Variables de punto de partida para identificar conglomerados

Variables	Precisiones
Número de cuartos en la vivienda	Excluye sanitarios y corredores.
Número de televisores	-
Número de computadoras	Computadora personal de escritorio, *lap top*, *ipad*.
Gasto en carne de ganado mayor	Res, cerdo, chivo, cordero.
Gasto en carne de pollo	-
Gastos en alimentos y bebidas fuera del hogar	-
Gastos en servicios del hogar	Servicio doméstico, lavandería, tintorería, jardinería, fumigación.
Gastos en cuidados personales	Corte de cabello y peinado, baños, masajes, etcétera.
Gastos en educación, cultura y recreación	Incluye, además de los educativos, gastos en entradas de cine, teatro, museos, exposiciones, conciertos y espectáculos deportivos.
Gastos en servicios de conservación de la vivienda	Además de mantenimiento, incluye cuotas de vigilancia, administración y recolección de basura.
Gastos diversos	Incluye, además de los gastos turísticos, paquetes para fiesta, seguros de automóvil y contra incendios.
Gastos en regalos otorgados	-
Gastos en luz y agua	-
Gastos en telefonía e Internet	Incluye la telefonía celular.
Pago de tenencia	-
Pago de tarjetas de crédito	-
Adquisición de activos	Erogaciones en el periodo de referencia (últimos seis meses) por compra de casas, condominios, locales o terrenos que no habita el hogar; compra de valores, cédulas acciones y/o bonos.

rentes (una cosa es una cantidad de computadoras y otra una magnitud monetaria) de modo que no se tiene algo sumable para trabajar únicamente con su sola magnitud agregada.

Pero la composición que hay detrás de las magnitudes importa asimismo, aun cuando todas las variables pudieran ser expresadas en una sola métrica común. En la estratificación que se emprende se buscan identificar conglomerados de hogares de acuerdo con su afinidad en términos de lo que poseen o gastan relativo a estas variables. En el límite, y para ilustrar el punto, si un $hogar_i$ gastara en uno solo de estos rubros mas no en los demás, pero alcanzando la misma magnitud monetaria de otros hogares que tienen un monto análogo repartido en todas las variables de gasto consideradas, ese $hogar_i$ en cuestión no quedaría en el mismo conglomerado de hogares afines: esto es una manera de decir que lo cuantitativo y lo cualitativo —o composición subyacente de lo cuantitativo— importan por igual.

Por lo demás, es común que cuando se realiza una selección inicial de múltiples variables se proceda a un análisis estadístico para detectar redundancias a partir de una marcada correlación entre algunas de ellas de modo que sea posible colapsar la selección primera en unos pocos indicadores de síntesis (componentes principales) o combinaciones lineales de esas 17 y que las sustituyan, todo ello para centrar la conglomeración, de ser posible, sólo en el componente o componentes que más contribuyan a la varianza detectada entre las observaciones (en este caso, los hogares).

Para esta investigación, resultó hasta cierto punto sorpresivo el comprobar la baja correla-

ción entre las variables de gasto seleccionadas. Ello obedeció, en buena medida, a que los casos de hogares con valores cero en esos rubros de gasto diferían de manera considerable de una variable a otra y, en general, a que las grandes diferencias entre hogares se marcan en distintos tramos de la distribución, según sea la variable.

Una manera de ilustrar lo anterior es con la gráfica 1, que representa las correlaciones entre variables donde el eje de las *X* es la primera componente principal y el de las *Y*, la segunda. Las flechas, por su parte, representan cada una de las variables, y el hecho de que tiendan a ángulos de 45 grados o cierta equidistancia con uno y otro eje nos habla de lo difícil que es sintetizar su información; ello también influye en el hecho de que la primera componente sólo dio cuenta de 21.8% de la varianza explicada, mientras que la segunda, de 9.1 por ciento.

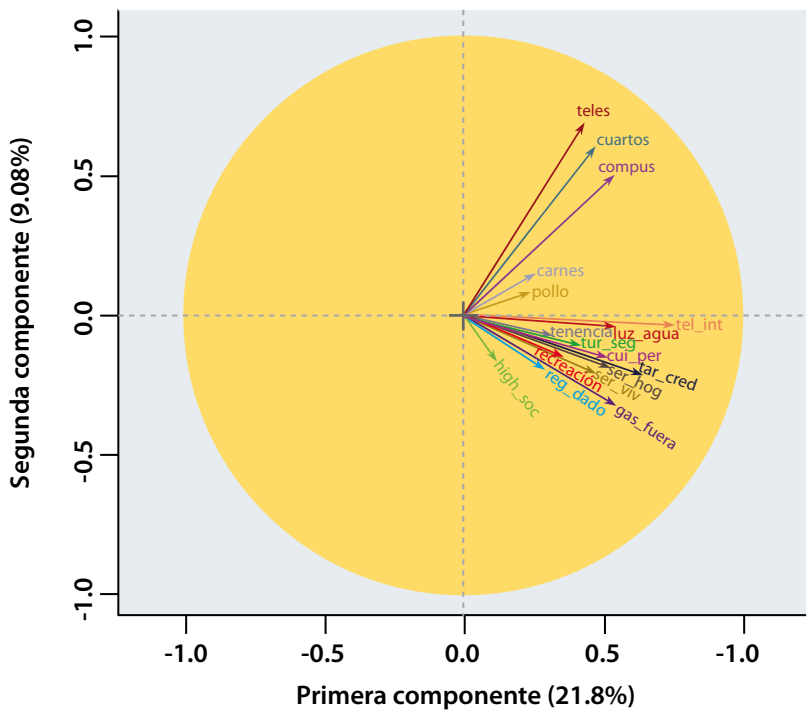
Dadas las 17 variables seleccionadas, se hubieran requerido de hasta 13 componentes principales para contribuir con más de 80% de la varianza explicada de las observaciones. Si hay tantos componentes como variables, no tiene caso colapsar a estas últimas ni perder información al hacerlo, porque básicamente cada una de ellas expresa algo que no dicen las demás: no otra cosa significa que haya una baja correlación entre ellas. De ahí que sea ineludible —desde un punto de vista estrictamente estadístico— proceder a una conglomeración fundamentada en un método multivariado.

3. Procedimiento seguido en esta investigación

Recapitulando lo dicho hasta ahora, en la exploración aquí emprendida se han privilegiado variables de gasto, así como otras que conectan con el acceso o posesión de activos por considerar que es un punto de confluencia de interés para economistas y sociólogos. A su vez, la estructura y naturaleza de estas variables nos indica que emprender un análisis multivariado es ineludible no sólo por quedar involucradas métricas diferentes sino, sobre todo, porque las variables en las que se ha fijado la aten-

Gráfica 1

Representación gráfica de variables en el plano de las primeras dos componentes principales



ción no se prestan a ser reducidas a una variable subyacente en esta fase inicial. Cabe observar, también, que todas estas variables de arranque son de carácter cuantitativo, no categórico.

El procedimiento que se emprende en esta investigación tiene aquí su punto de partida: elegir un método que permita identificar el mejor modelo de agrupamiento o conglomeración de los hogares observados en la ENIGH 2010 tomando en cuenta estas variables de arranque. La decisión del modelo a seguir es una fase netamente algorítmica; sin embargo, ya que se tienen establecidos los conglomerados o grupos homogéneos de hogares, la naturaleza del problema —identificación de clases sociales— obliga tanto a ordenarlos como a establecer a partir de qué conglomerado se detecta una diferencia sustancial con los que le anteceden en el ordenamiento. Ésta es una fase analítica en la que interviene no sólo lo cuantitativo, sino también variables categóricas que resulta posible introducir una vez que la complejidad del problema inicial se redujo en virtud de la etapa previa de conglomeración. Cuando se ha hecho esto, se puede aventurar en dónde terminan grupos de afinidad atribuibles a una clase social y a partir de qué grupos de afinidad comienza otra (ver tabla 2).

3.1 Conglomeración por el método de selección de modelos

Explorar por este método significa evaluar diferentes formas de conglomeración con distintas geometrías o parametrizaciones —definidas por las características de la matriz de covarianza en cada caso— y seleccionar de manera específica el modelo de conglomeración que, dado lo que se observa en la muestra ENIGH 2010, sea el más factible de desprenderse del universo poblacional que se ha decidido describir a partir de 17 variables y del cual se infieren sus propiedades desde dicha muestra.

El significado matemático de la estimación del conjunto de parámetros que subyacen a un modelo (a partir del concepto general de una función de mezcla de distribuciones o distribución conjunta), así como el de una selección final del modelo de conglomeración más factible de corresponder al universo poblacional a partir de la optimización de lo que se conoce como *Bayesian Information Criteria* (BIC), se describe con amplitud en el *Anexo*. En lo que sigue, se tratará de explicar de manera intuitiva qué queremos decir con conglomerar con geometrías o plantillas de conglomeración y, asimismo, en qué momento señalamos que no presuponemos una geometría o configuración

Tabla 2

Tres fases seguidas en el presente ejercicio de identificación y cuantificación de la clase media en México

Fase	Objetivo	Método	Unidad de observación	Variables
Primera	Identificar grupos homogéneos de hogares y dejar que el algoritmo determine cuáles y cuántos son	Método de selección de modelos para la conglomeración de hogares	Hogares	Variables cuantitativas
Segunda	Ordenar los grupos homogéneos y ubicar a partir de cuál conglomerado se detecta una diferencia manifiesta con los conglomerados que le preceden	Método de *Dalenius-Hodges* y componentes principales	Conglomerados	Variable cuantitativa y variables categóricas
Tercera	Formar las clases sociales utilizando los conglomerados como sus bloques de construcción y cuantificarlas	Conclusión del analista a partir de las evidencias proporcionadas por las fases previas	Conglomerados	No intervienen más variables

del conglomerado o grado de complejidad del mismo, error este último que es el más frecuente entre quienes optan por tal o cual estratificación multivariada sin estar al tanto de que pueden estar imponiendo una geometría de conglomeración sobre una realidad configurada de otra forma.

Los hogares en la base de datos de la ENIGH 2010 son las observaciones a clasificar en grupos o conglomerados de afinidad. Si, por ejemplo, éstas (en este caso los hogares) se clasificaran sólo en función de dos variables, cada observación tendría dos coordenadas, *X* y *Y*, que definen en un plano cartesiano cada grupo bidimensional de afinidad representado por manchas en la gráfica 2.

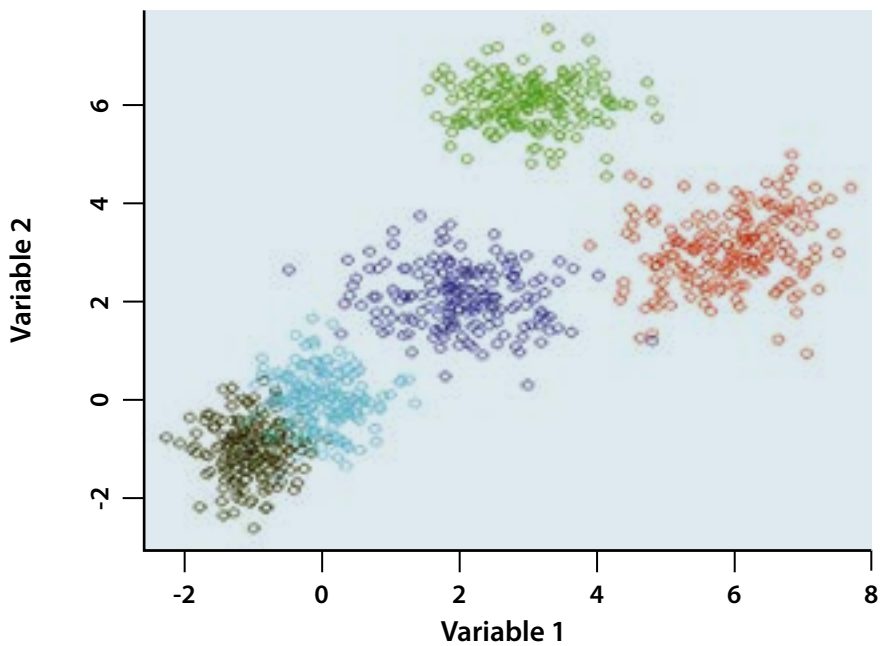
Ahora bien, si tomamos en cuenta ya no dos variables para clasificar, sino $n = 3$ para hacerlo, pasamos entonces a un espacio donde aquellos patrones de manchas indicativos de la existencia de grupos distintos ahora adquieren un volumen a manera de nubes; pero si vamos más lejos todavía e involucramos 17 variables, cabe concebir —aunque sea imposible visualizar— esa configuración análoga de lo que serían nubes o conglomerados en el espacio tridimensional. Hablaríamos ya no de nubes, sino de híper nubes, donde cada una agrupa hogares afines en el espacio así estructurado desde 17 ejes.

Una vez que entramos a ese nivel de complejidad, más allá de la visualización, cabe pensar que, para encontrar las híper nubes que subyacen a una dispersión de observaciones con *m* coordenadas en un espacio *n* dimensional, hay que introducir hipótesis sobre qué geometrías o configuraciones pueden tener, y así dar con el patrón que mejor se ajusta al aparente caos de datos, llevándonos a una óptima identificación de grupos. Dicho de otra forma, la clave del método basado en modelos es que no prejuzga cómo están configurados los cúmulos de hogares: lo que hace es poner a prueba distintas plantillas de cúmulos o híper nubes para poder descifrar qué configuración se ajusta mejor a lo que se observa.

Cada plantilla o patrón es un modelo que puede tolerar o no variaciones entre las híper nubes que comprende. Las plantillas más básicas, como la de *K-medias*, presuponen que cada nube es igual a la otra en forma y orientación, y que los datos están agrupados en esferas que se distribuyen a lo largo

Gráfica 2

Población en dos dimensiones (cinco grupos)



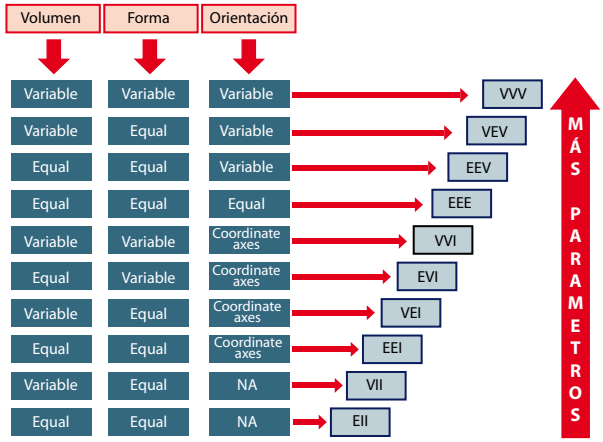
del plano *n* dimensional; otros modelos permiten más complejidad: que las nubes, además de variar en volumen, también se orienten de diferente manera y hasta cada una puede tener formas o geometrías distintas; asimismo, que no necesariamente están separadas, sino que puede haber tramos de superposiciones (ver figura 1).

Las propiedades de las híper nubes requieren que sean especificados ciertos parámetros (mientras más compleja la configuración de nubes, más parámetros se requieren) mismos que se deben obtener a partir de la estructura de las matrices de covarianza de los datos con los que se ha decidido trabajar (ver figura 2).

Estos parámetros se estiman con un método iterativo bajo el principio de máxima verosimilitud, y con esos estimadores se calcula, a su vez, la probabilidad de que la distribución de hogares de cada modelo en un espacio *n* dimensional que se observa en la muestra ENIGH 2010 se haya desprendido del universo poblacional. Una forma de detectar qué modelo o modelos mejor califican por ser los más factibles de provenir de la configuración poblacional es visualizando el(los) que más alto se sitúa(n) en una escala transformada[5] que se desprende del

Figura 2

Parámetros que caracterizan a los distintos modelos de conglomeración



5 El criterio BIC busca el valor negativo mayor como criterio de optimalidad para seleccionar modelos (el significado de ello se explica en el Anexo); sin embargo, para ilustrarlo, se invierte el signo de modo que los modelos seleccionados aparecen en el tramo superior del gráfico.

Figura 1

¿Cuál es el patrón o plantilla de cúmulos que subyace tras la dispersión de observaciones en un espacio *n* dimensional?

denominado criterio de información de *Bayes*, también conocido como BIC (ver el *Anexo*).

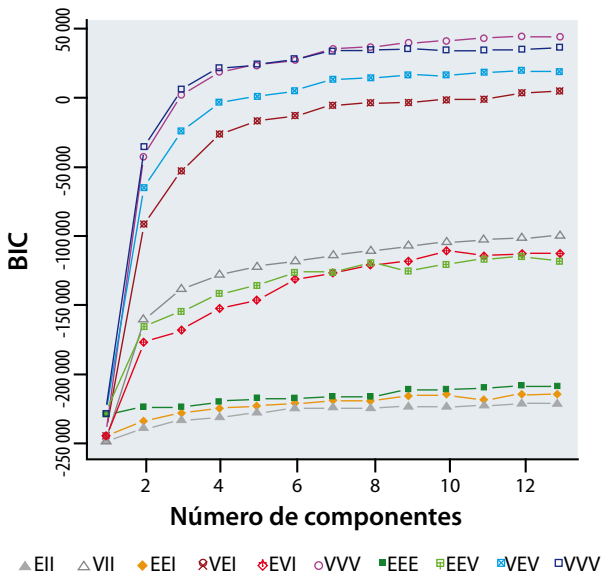
En la gráfica 3, cada línea de color corresponde a un tipo de modelo; en el eje de las *X* se representa el número de conglomerados y en el de las *Y*, la escala BIC. De los modelos sujetos a prueba en este ejemplo, los que tienen las probabilidades más altas de describir la configuración subyacente de conglomerados en el universo poblacional son los VVI y VVV, que fueron los obtenidos respectivamente para los dominios urbano y rural de la ENIGH 2010.[6]

El número óptimo de conglomerados se toma a partir del momento en que el avance en la escala BIC se torna marginal. Tanto en el caso del dominio urbano como en el del rural de la ENIGH 2010 se consideró que ese punto se alcanza con siete conglomerados. Una vez seleccionado el modelo y el número de ellos, se adopta su función de distribución conjunta, misma que determina las probabilidades de pertenencia de cada hogar en la base de datos ENIGH a cualquiera de los cúmulos ya especificados por esa función y los va asignando de acuerdo con la mayor probabilidad de pertenencia al cúmulo o conglomerado correspondiente.

6 Para efectos de este análisis, el domino urbano en la ENIGH 2010 es el de 15 mil o más habitantes y, por ende, el rural es el de menos de 15 mil. Los modelos fueron probados por separado para uno y otro dominio.

Gráfica 3

Modelos en la escala transformada BIC



Todo lo anterior no significa que una vez que seleccionemos la plantilla o configuración de nubes —el modelo que mejor se ajusta a los datos— en automático arrojará cuáles son las clases medias en el país, pero sí al proporcionar un número determinado de conglomerados nos facilita el análisis para saber cuáles de ellos se aproximarán o no a una noción de clases medias. Para el método estadístico, hablando de manera estricta, no existe el término clase media: al identificar un cierto número de conglomerados, simplemente nos dice que éstos son los que mejor describen los grupos de afinidad que subyacen a la dispersión de las observaciones en el espacio *n* dimensional, dadas las variables seleccionadas.

La conglomeración multivariada nos ofrece un primer resumen en nuestra tarea de investigación que luego habrá de afinarse con las denominadas variables informativas o de análisis: promedios que se desprenden de la conformación de los conglomerados, pero que en sí mismas no intervienen en formarlos. Estas variables informativas o analíticas son ya un resultado de dicha conglomeración y nos guiarán hasta nuestra meta final.

La originalidad de la exploración que aquí se adopta consiste, entonces, en la conjunción de cuatro características: a) una aproximación al fenómeno por la vía de una metodología de conglomeración, b) que ésta es multivariada, c) que se adopta la conglomeración multivariada después de someter a pruebas estadísticas de verosimilitud a distintos modelos y d) el análisis subsecuente de los resultados de la conglomeración que, de antemano, no se sabe o no se prejuzga en qué han de desembocar.

3.2 Ordenamiento y análisis de la conglomeración obtenida

Tenemos, así, 14 conglomerados (siete en el ámbito urbano y siete en el rural) o grupos de afinidad

de hogares en función de las variables utilizadas. Tomando ahora el gasto o las erogaciones de éstos como hilo conductor, ¿cómo quedan ordenados y cuáles son las brechas que quedan definidas entre ellos?, ¿en qué desemboca pues la conglomeración vista bajo esa perspectiva?

Para introducir un orden en esos 14 conglomerados multivariados y tener un primer acercamiento de la jerarquía entre ellos, se puede ensayar una estratificación de *Dalenius* con base en el promedio del gasto corriente per cápita total (neto de gastos funerarios y hospitalarios) que arroja o se desprende de cada conglomerado. Vale la pena recalcar que de aquí en adelante hablaremos de *conglomerado* (grupos de afinidad) cuando tenemos el resultado bruto del procedimiento multivariado y de *estrato* cuando introducimos un orden entre ellos en función de una variable: una, además, que como tal no estaba implicada en la conglomeración (sólo lo estaban ciertos rubros de gasto).[7]

Los conglomerados se ordenan, entonces, en función del gasto corriente per cápita y se les asigna un numeral arábigo descendente donde el 7 corresponderá al valor más alto en su ámbito y la letra designa si éste es urbano (*u*) o rural (*r*).

En la tabla 3 se observan sucesivos reagrupamientos de los conglomerados con la variable de gasto corriente per cápita con $n = 3, 4, 5, 6$ ó 7 estratos. En cada columna se muestra con colores como quedan reagrupados los conglomerados originales (numeral arábigo) en bloques más grandes o estratos (numeral romano). Así, por ejemplo, cuando los conglomerados se colapsan en cinco estratos (columna $n = 5$) se tiene que el conglomerado 7 urbano queda aislado de todos y se convierte en el estrato I; los conglomerados 6 urbano y 7 rural se colapsan en el estrato II; el 5 y el 4 urbanos quedan asociados como estrato III, mientras que el conglomerado 3 urbano en combinación con los 5 y 6 rurales forman otro estrato y, por último, como estrato V, con los valores menores, queda el bloque formado por el resto de los conglomerados.

Los sucesivos ejercicios mandan una primera señal sistemática: el conglomerado 7 urbano nunca queda asimilado a ningún otro: siempre será el estrato I, es decir, ninguno tiene cercanía a él en términos de nivel de vida; simplemente *se cuece aparte*. Por otro lado, los conglomerados 6 urbano y 7 rural en cuatro de cinco estratificaciones quedaron asociados. En el polo opuesto están los tres primeros rurales que siempre van juntos con una tendencia a aislarse de los demás. Un caso en particular interesante para los fines del análisis que nos ocupa lo da el conglomerado 3 urbano con su tendencia a separarse de los dos más bajos de su ámbito por una parte y, por la otra, a nunca romper su asociación con los conglomerados 6 y 5 rurales (en ningún caso quedan en estratos distintos). Ahora bien, ¿por qué esto último importa en esta fase del análisis?

La razón es que si ese estrato urbano no sólo no está en el fondo de su ámbito, sino que su tendencia en función del gasto corriente per cápita es marcar una distancia importante con los dos primeros estratos urbanos que, por su parte, sí son sistemáticos en quedar juntos, ello es una pri-

7 Cabría preguntar por qué si se está utilizando una estratificación univariada (*Dalenius*) tomando como referencia el gasto corriente en esta fase del análisis no se hizo desde el principio y así evitar más complicaciones en la metodología seguida. Por ello, es importante reiterar que los conglomerados no sólo se constituyeron en función de algo meramente cuantitativo, sino a partir de afinidades en la estructura de gasto corriente centrándonos en ciertos rubros de éste. Si el punto de partida hubiera sido una estratificación univariada que sólo sigue la magnitud total del gasto, dicha magnitud, por sí misma, lo diría todo. La diferencia entre grupos podría ser de centavos sin otro fundamento; punto. Pero la estructura compleja que resultó del tipo de conglomerado más verosímil implicaría que, en cuanto a la magnitud de gasto total (o cualquier otra variable cuantitativa única), pudiera haber traslapes entre los conglomerados. Después de todo, en esta fase —posterior a la conglomeración de hogares— estamos tratando vía *Dalenius* con los promedios de gasto per cápita de los conglomerados, pero pudiera haber observaciones (hogares) o subgrupos de observaciones dentro de un mismo conglomerado_k con niveles de gasto per cápita inferiores a algunos otros hogares situados en un conglomerado_j próximo que, como tal, tuviera un promedio inferior al del conglomerado_k de pertenencia. De ahí que la sola magnitud cuantitativa no fue determinante. Así, la estratificación de *Dalenius* se utiliza aquí como un expediente de ordenación de conglomerados que le precede más no de los hogares en sí mismos; si el ordenamiento funciona es porque el conjunto de rubros específicos de gasto seleccionados, en la fase multivariada, tienen una buena correlación con el gasto total pero introduciendo matices que un procedimiento univariado por sí solo no detecta en la conformación de afinidades entre los hogares. Por lo demás, utilizar una estratificación univariada (sea con el total del gasto o del ingreso corriente) como único referente para agrupar hogares desde el punto de arranque no daría una sola pista de cuántos estratos utilizar o cuál sería el número óptimo, más allá de la solución trivial en la que el número de estratos sea igual al número de observaciones (suma de la varianza de los estratos = 0). El método seguido de conglomeración, en cambio, sí permite una decisión fundamentada del número óptimo de grupos de afinidad. La filosofía adoptada es que no hay que eludir la complejidad; sólo una vez que la hemos abordado se puede despejar el camino para que los procedimientos más simples puedan ser adoptados de una manera analíticamente más eficiente y segura.

Tabla 3

Resultado del ordenamiento de conglomerados por el método de *Dalenius* en función de la variable de gasto

Conglomerados ordenados como estratos	Gasto mensual per cápita	*n* = 3	*n* = 4	*n* = 5	*n* = 6	*n* = 7	Clasificación tentativa
		Colapsados en tres estratos	Colapsados en cuatro estratos	Colapsados en cinco estratos	Colapsados en seis estratos	Colapsados en siete estratos	
7*u*	15 617	I	I	I	I	I	Alta absoluta
6*u*	8 317	II	II	II	II	II	Media alta
5*u*	5 307	II	II	III	III	IV	Media media
4*u*	4 008	III	III	III	III	IV	Media media
3*u*	2 890	III	III	IV	IV	V	Media baja
2*u*	2 114	III	IV	V	V	VI	Baja
1*u*	1 821	III	IV	V	V	VI	Baja
7*r*	6 704	II	II	II	II	III	Media alta
6*r*	2 755	III	III	IV	IV	V	Media baja
5*r*	2 500	III	III	IV	IV	V	Media baja
4*r*	1 578	III	IV	V	V	VI	Baja
3*r*	1 431	III	IV	V	VI	VII	Baja absoluta
2*r*	1 228	III	IV	V	VI	VII	Baja absoluta
1*r*	958	III	IV	V	VI	VII	Baja absoluta

mera señal indicativa de que ahí es donde pudiera iniciar la frontera de una clase media. A su vez, y por su tendencia a la asociación con los conglomerados 6 y 5 rurales, podríamos inferir dónde inicia la clase media rural ya que los dos conglomerados urbanos más bajos y el 4 rural tienden a constituirse como un bloque aparte.

Esta manera de analizar la conglomeración es muy sugestiva, mas no decisiva. Se necesitan corroborar otros aspectos, tomar en cuenta características en los conglomerados en función de variables que no predeterminaron la conglomeración, pero que, en cambio, nos pueden decir en qué desembocó ésta en términos de promedios relativos a nivel de instrucción, presencia de trabajo intelectual, calificación y jerarquía ocupacional, así como acceso a la propiedad, por mencionar algunas de relevancia sociológica.[8]

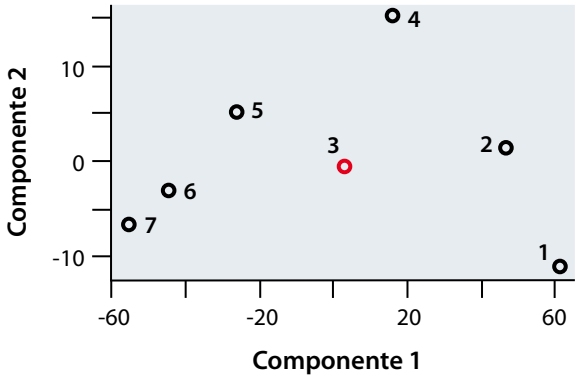
Una peculiaridad al jugar con estas variables que promedian lo que ahora serían las observaciones (en este caso los conglomerados urbanos) es que fácilmente pueden sintetizarse en sólo dos componentes principales o combinaciones lineales de las mismas, siendo capaces de dar cuenta de 98% de la varianza explicada. Es así que al graficar los siete conglomerados urbanos en un plano cartesiano

8 Las variables en cuestión utilizadas para analizar los siete conglomerados urbanos son: a) hogares cuyo jefe tiene estudios medio superiores o superiores, b) hogares con vivienda propia escriturada, c) hogares con algún ocupado en trabajo intelectual no directivo, d) hogares con algún ocupado en trabajos manuales operativos no calificados y e) hogares con algún integrante en pobreza conforme a la medición multidimensional de la misma diseñada por el Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL).

—en el eje de las *X* se encuentra la primera combinación lineal o primera componente, mientras que en el de las Y, la segunda—, el 3 urbano claramente se sitúa a medio camino en ambos ejes, lo cual lo confirma como un referente clave (ver gráfica 4).

Gráfica 4

Ubicación de conglomerados urbanos en el plano de los componentes principales



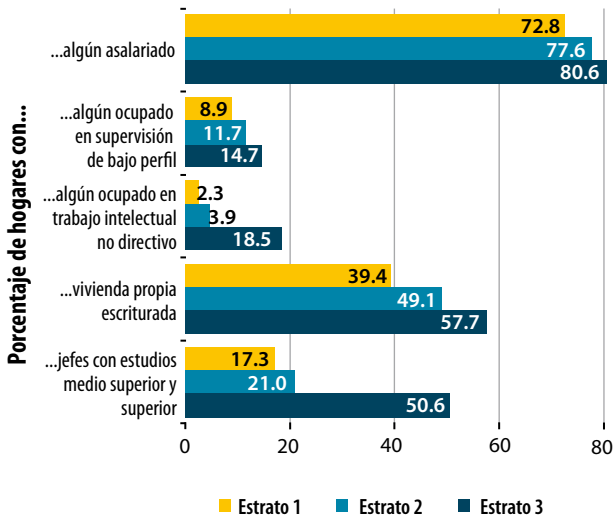
Notas: a) las primeras dos componentes explican 98% de la varianza y b) las variables que se usaron para la agrupación (en términos de porcentaje de cada una son: hogares cuyo jefe tiene estudios medio superiores o superiores, hogares con vivienda propia escriturada, hogares con algún ocupado en trabajo intelectual no directivo, hogares con algún ocupado en trabajos manuales-operativos no calificados y hogares con algún pobre bajo la metodología multidimensional.

Basta comparar simplemente los promedios del conglomerado 3 con los dos que le anteceden para darnos cuenta de su diferencia en cuanto a estas variables. Dado entonces su nivel de inserción en el mercado de trabajo y su mayor nivel de calificación, es palpable que refleja un nivel de vida distinto de los conglomerados que le preceden, marcando una distancia cuantitativa y cualitativa (ver gráfica 5).

Es importante recalcar que no se está diciendo que el conglomerado 3 sea, por sí solo, el de la clase media urbana sino, más bien, que es ahí donde ésta comienza. Las diferencias con conglomerados superiores no necesariamente entrañan una diferencia de clase; por ejemplo, el conglomerado 3 presenta promedios de instrucción e inserción laboral superiores al 4 aun cuando su nivel de gasto per cápita sea inferior. En este

Gráfica 5

Comparativo del conglomerado 3 urbano con los que le preceden



tipo de diferencias no hay que perder de vista los factores demográficos o composiciones diferentes en términos de ciclo de vida de los hogares. En nuestro análisis no dejó de llamar la atención que el promedio de edad del jefe del hogar o el de los integrantes en su conjunto claramente se sitúa en el conglomerado 3 por debajo del 4 (ver tabla 4). Detrás de esos promedios puede haber mayor preponderancia de hogares en el 4 que ya iniciaron su proceso de fisión (hijos que salen a poner su propio hogar), así como el hecho de que esté cristalizando como capital humano una mayor experiencia entre quienes están insertos de hace tiempo en el mercado de trabajo o, también, se tenga una mayor presencia de hogares beneficiados de algún tipo de renta de la propiedad o transferencia (remesas) que son fuentes de ingreso por lo normal vedadas a los hogares más jóvenes o recientes.

Esta relación entre el conglomerado 3 y el 4 la volvemos a encontrar de algún modo entre los 5 y 6, pues este último, si bien presenta un gasto per cápita claramente superior, también no deja de resaltar la diferencia de promedios de edades como un factor subyacente a tomar en cuenta, factor que, a su vez, no puede abonar a la explicación de por qué el conglomerado 7 tiene un nivel de gasto

Tabla 4

Comparativo de conglomerados urbanos según edades promedio de sus integrantes

Conglomerados ordenados como estratos	Edad promedio del jefe del hogar	Edad promedio de todos los integrantes del hogar
7	49.2	36.9
6	50.9	36.1
5	46.7	31.5
4	51.2	36.2
3	47.5	29.5
2	48.7	28.9
1	47.1	28.3

per cápita superior en 87.8% al del 6, la mayor brecha al respecto entre conglomerados vecinos en el ámbito urbano.

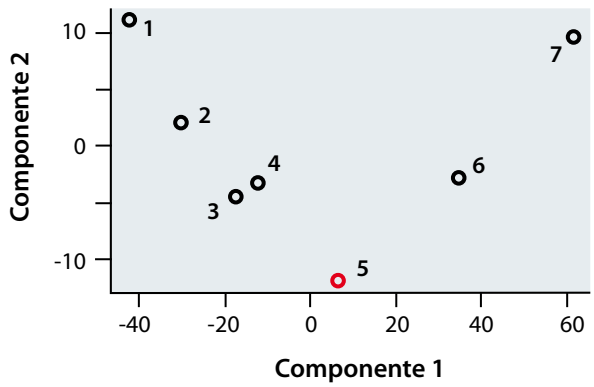
Por todo lo anterior, una primera conclusión a la que se arriba en este análisis es que los hogares comprendidos en los conglomerados 3, 4, 5 y 6 forman el subuniverso de la clase media urbana con sus distintos matices.

¿A qué nos conduce el análisis de conglomerados en el ámbito rural? Si bien en éste no tiene sentido tomar con exactitud las mismas variables cualitativas que se utilizaron en el ámbito urbano (dado que la jerarquía ocupacional está menos desarrollada o el tener una vivienda propia que poco discrimina en esos ámbitos), sí cabe incorporar un conjunto de variables que, de cualquier forma, nos hablan de estatus, posición y tipo de inserción laboral[9] para realizar un análisis de componentes que identifique el conglomerado a partir del cual se marca una clara diferencia (ver gráfica 6).

Aquí, también, el colapso de las variables cualitativas en dos componentes da cuenta de 98% de la varianza explicada entre los conglomerados y al utilizar la métrica de ambos componentes como ejes en un plano cartesiano resalta el punto de quiebre que representa el conglomerado 5.[10]

Gráfica 6

Distribución de conglomerados rurales en el plano de los componentes principales



Notas: a) las primeras dos componentes explican 98% de la varianza y b) las variables que se usaron para la agrupación, en términos de porcentaje de cada una, son: hogares cuyo jefe tiene estudios medios superiores o superiores, hogares con algún dueño de micronegocio no agropecuario, hogares con algún asalariado agropecuario, hogares con algún ocupado en trabajos de supervisión, hogares con algún ocupado en trabajos manuales u operativos no calificados y hogares con algún pobre bajo la metodología multidimensional diseñada por el CONEVAL.

9 Las variables cualitativas utilizadas para el ámbito rural son: a) hogares cuyo jefe tiene estudios medios superiores o superiores, b) hogares con algún dueño de micronegocio no agropecuario, c) hogares con algún asalariado agropecuario, d) hogares con algún ocupado en trabajos de supervisión, e) hogares con algún ocupado en trabajos manuales u operativos no calificados y f) hogares con algún pobre bajo la metodología multidimensional diseñada por el CONEVAL.

10 Cabe observar que las variables utilizadas en el ámbito rural (ver nota 6) ponen más énfasis en encontrar diferencias en la parte baja del estatus o jerarquía ocupacional que en la alta (ámbito urbano) porque es más difícil esa última diversificación en el rural. Esto, a su vez, hace que en la gráfica de componentes el orden de los conglomerados aparezca invertido con respecto al urbano por el alejamiento, respecto al origen, de los conglomerados de mejor nivel en lugar de cercanía. Ello no afecta el análisis porque el propósito aquí es ver si hay un conglomerado *parteaguas*, no importando el sentido del ordenamiento.

En efecto, es ese mismo conglomerado que en la estratificación de *Dalenius,* en términos de la variable de gasto per cápita, lo dejaba asociado al 3 urbano.

La conclusión es que, efectivamente, la clase media en el ámbito rural comienza a partir del conglomerado 5. Por lo demás, basta mirar la distancia que guarda el 7 urbano con respecto al 7 rural en términos de gasto per cápita (superándolo casi en 133%) para comprender por qué la estratificación de *Dalenius-Hodges* asocia al séptimo rural con el conglomerado 6 urbano en cinco de seis estratificaciones y, en la que no, queda incluso por debajo de este último; en otras palabras, en el ámbito rural no hay una clase alta propiamente dicha, pero sí una clase media que comprende tres conglomerados de hogares: el 5, 6 y 7.

3.3 Conformación de las clases sociales a partir de los conglomerados y cuantificación de las mismas

Tenemos, entonces, los conglomerados ordenados y sabemos, además —por el análisis precedente—, que en el ámbito urbano, la clase media queda comprendida entre el estrato 3 y el 6 mientras que, en el rural, abarca sus tres conglomerados superiores. Con estas referencias, podemos ver cómo queda configurado el panorama de las clases sociales en México.

La tabla 5 muestra, por medio de áreas de color, los conglomerados que quedan involucrados en cada clase social junto con las magnitudes de hogares y población que cada uno por separado aporta en sus respectivos ámbitos.

Tabla 5

Correspondencia de conglomerados a clases sociales

Conglomerados ordenados como estratos	Hogares en el ámbito urbano	%	Población en el ámbito urbano	%	Hogares en el ámbito rural	%	Población en el ámbito rural	%
	18 821 246	100.00	70 284 500	100.00	10 146 938	100.00	42 007 455	100.00
7	724 689	3.85	1 919 539	2.73	377 937	3.72	1 132 954	2.70
6	1 782 504	9.47	5 249 790	7.47	1 415 500	13.95	6 039 552	14.38
5	2 409 112	12.80	8 512 475	12.11	1 060 149	10.45	3 759 264	8.95
4	2 109 223	11.21	5 652 475	8.04	2 028 387	19.99	9 028 885	21.49
3	3 134 215	16.65	13 623 582	19.38	772 590	7.61	3 570 814	8.50
2	3 526 126	18.73	15 612 159	22.21	3 599 016	35.47	14 969 389	35.64
1	5 135 377	27.29	19 714 480	28.05	893 359	8.80	3 506 597	8.35

Clase alta | Clase media | Clase baja

Acumulando los hogares y personas de cada bloque de color se tienen, entonces, las respectivas magnitudes totales para clase social tanto en su propio ámbito como a nivel nacional.

Al no haber tal cosa como una clase alta en zonas rurales (localidades de menos de 15 mil habitantes), este segmento —únicamente presente en el ámbito urbano a nivel nacional— representa 2.5% de los hogares y 1.7% de la población; la clase media, por su parte, comprende a nivel nacional 42.4% de los hogares e involucra 39.2% de las personas, alcanzando la mayoría de los hogares mas no de los residentes en el ámbito urbano. En cuanto a la clase baja resulta la mayoría de los hogares y población al término de la primera década del siglo XXI (ver tabla 6).

Una vez confirmado el análisis que nos precisa en qué conglomerado comienza la clase media en cada ámbito y dónde termina, podemos retomar con confianza la subclasificación de la clase media que se desprende del ordenamiento *Dalenius* (ver tabla 7) para cuantificar, a su vez, cómo podría ser la composición interna de dicho segmento social.

Es así que de los casi 12.3 millones de hogares de clase media, 45.7% calificarían como media baja aportando 53.3% de población de la clase, 36.8% de los hogares y 32.2% de la población sería la parte intermedia, mientras que 17.6% de los hogares con 14.5% de población del segmento social corresponderían a la clase media alta a nivel nacional. Es asimismo interesante constatar una diferencia notable del ámbito urbano con respecto al rural

Tabla 6

Magnitud de las clases sociales en México en hogares y población

Clase	Nacional				Urbano				Rural			
	Hogares	%	Población	%	Hogares	%	Población	%	Hogares	%	Población	%
Alta	724 689	2.5	1 919 539	1.7	724 689	3.9	1 919 539	2.7	_	_	_	_
Media	12 288 640	42.4	43 970 092	39.2	9 435 054	50.1	33 038 322	47.0	2 853 586	28.1	10 931 770	26.0
Baja	15 954 855	55.1	66 402 324	59.1	8 661 503	46.0	35 326 639	50.3	7 293 352	71.9	31 075 685	74.0
Total	28 968 184	100.0	112 291 955	100.0	18 821 246	100.0	70 284 500	100.0	10 146 938	100.0	42 007 455	100.0

Tabla 7

Segmentación de la clase media en México

Niveles y conglomerados	Nacional				Urbano				Rural			
	Hogares	%	Población	%	Hogares	%	Población	%	Hogares	%	Población	%
Media alta (6 urbano; 7 rural)	2 160 441	17.6	6 382 744	14.5	1 782 504	18.9	5 249 790	15.9	377 937	13.2	1 132 954	10.4
Media media (4 y 5 urbano; ninguno rural)	4 518 335	36.8	14 164 950	32.2	4 518 335	47.9	14 164 950	42.9	0.00	0.0	0.0	0.0
Media baja (3 urbano; 5 y 6 rural)	5 609 864	45.7	23 422 398	53.3	3 134 215	33.2	13 623 582	41.2	2 475 649	86.8	9 798 816	89.6
Total	**12 288 640**	**100.0**	**43 970 092**	**100.0**	**9 435 054**	**100.0**	**33 038 322**	**100.0**	**2 853 586**	**100.0**	**10 931 770**	**100.0**

pues, en el primero, su segmento intermedio es el predominante de los tres en términos de hogares y ligeramente en cuanto a población, mientras que en el ámbito rural no hay tal segmento intermedio de clase media: no hay ahí un gradiente entre las clases media baja y alta.

3.4 De pobres y ricos

Más allá de lo que aquí se identifica como clase media, es importante subrayar que en esta investigación se utiliza de manera deliberada el término clase baja y no pobre, ya que esto último corresponde a una definición precisa del CONEVAL como una combinación de deficiencias de ingresos corrientes y de carencias en términos de garantías o derechos sociales.[11] Lo que aquí se establece, entonces, es que no necesariamente todo aquél en clase baja está en pobreza en el sentido de quedar ubicado debajo de un umbral normativo de ingresos y de acceso a bienes y servicios públicos que impida ejercer sus capacidades básicas como miembros de la colectividad nacional. Del mismo modo que hay individuos marginados de los mercados de trabajo, así como individuos depauperados, hay familias de trabajadores (de cuello azul) no pobres vinculados a los mecanismos de seguridad social, protección al trabajo y acceso a los bienes públicos y cuyos miembros están en posición de ejercer sus facultades ciudadanas.

En consecuencia de este estudio exploratorio se desprende que la pobreza más que constituir una clase social en sí misma es una condición que puede presentarse con mayor probabilidad para un segmento que corresponde a 55.1% de los hogares y 59.1% de la población del país.[12] Eventos catastróficos al interior del hogar, como la pérdida súbita del principal proveedor o la presencia de una enfermedad o accidente grave entre sus integrantes, pueden ser factores decisivos para que ese segmento incurra en pobreza, lo mismo que un episodio de hiperinflación o una recesión profunda en el plano macroeconómico. Por ello, la condición de pobreza fluctúa más que la pertenencia a una clase social propiamente dicha. Así, la clase baja (lo mismo que la media) resulta un segmento heterogéneo, pero estable, en el que se presentan distintas situaciones de previsión frente a la adversidad, de cercanía a los mecanismos de protección al Estado y de pertenencia a redes de solidaridad grupal.

Por tener tanto los datos del CONEVAL como la exploración aquí seguida la misma fuente (en este caso, la ENIGH 2010), cabe hacer una comparación directa entre las mediciones de pobreza y clase baja en el país.

De lo anterior se desprende que al menos 13.6 millones de personas clasificadas en clase baja no son pobres, esto es, alrededor de una quinta parte de ésta, aunque no por ello afines a la clase media (ver tabla 8). Esto no es irrelevante: es, de hecho, un subsegmento fuera de foco en un debate que sólo habla de pobres o clase media, pero considerablemente menos del ámbito del trabajo, omisión que no ha dejado de ser muy sintomática del discurso y la esfera pública casi desde el inicio del proceso de la alternancia democrática en México. Queda claro que una diferencia que no se debe perder de vista de la presente exploración es que, en ningún momento, dio por hecho que ahí donde termina la pobreza comienza una clase media.

Por lo demás, si bien esta exploración por el reagrupamiento de conglomerados vía estratificación de *Dalenius* pareciera sugerir que existen

11 CONEVAL, en su sitio en internet (http://www.coneval.gob.mx/Medicion/Paginas/Medici%C3%B3n/Que-es-la-medicion-multidimensional-de-la-pobreza.aspx), remite al siguiente texto: "La población en situación de pobreza será aquélla cuyos ingresos sean insuficientes para adquirir los bienes y servicios que requiere para satisfacer sus necesidades y presente carencia en al menos uno de los siguientes indicadores: rezago educativo, acceso a los servicios de salud, acceso a la seguridad social, calidad y espacios de la vivienda, servicios básicos en la vivienda, servicios físicos en la vivienda y acceso a la alimentación…". CONEVAL. "Lineamientos y criterios generales para la definición, identificación y medición de la pobreza", en: *Diario Oficial de la Federación*. Título segundo, párrafo octavo, 16 de junio de 2010, pp. 12.

12 Es interesante, por lo demás, conectar esta conclusión con la discusión que plantea el sociólogo norteamericano Jack Metzgar sobre si el concepto de clase de pobres ha sustituido o no al clásico de clase trabajadora (*blue collar*) y sus implicancias en el debate público en Estados Unidos de América con respecto al estado de bienestar, sus críticos y detractores, ver Metzgar, J. "Are the Poor Part of the Working Class or in a Class by Themselves?", en: *Labor Studies Journal*. Vol. 35, Núm. 3, septiembre de 2010.

Tabla 8

Comparativo de personas en clase baja y pobreza en el 2010

	% de la población total	Monto absoluto (miles de personas)
Clase baja	55.1	66 402
Pobres CONEVAL	46.1	52 813
Diferencia	9.0	13 589

dos subconjuntos irreductibles al interior de la clase baja, siendo uno de ellos netamente rural (ver tabla 3), no se alcanza a percibir un matiz similar en lo que a la alta se refiere dado que se detecta un único conglomerado que se aleja del resto de una manera casi exponencial (el 7 urbano). El que ello sea así puede tener relación con el llamado problema de truncamiento que enfrentan las encuestas por muestreo probabilístico de hogares no sólo en México, sino en el resto del mundo, el cual se refiere a la bajísima probabilidad de caer en muestra por parte de quienes poseen las mayores fortunas, amén del problema que se desprende de no conceder una entrevista tipo ENIGH al operativo de campo de la encuesta por cuestiones de seguridad. A lo anterior se suma el comprometer una confidencialidad estadística que, fácilmente, se nulifica cuando se tiene un conjunto muy localizado de observaciones extremas, pues quedan en evidencia en la base de datos los hogares que más riqueza concentran del país si se les asigna de manera predeterminada una probabilidad 1 (total certeza) de aparecer en muestra.[13]

Dado ese tipo de truncamiento,[14] es posible que quedara fuera y por arriba por lo menos un conglomerado más de hogares o varios conglomerados, pero con un número bajísimo de hogares cada uno porque el proceso de diferenciación entre lo remanente es aún más extremo que en lo observado y con distancias progresivamente mayores con lo que le precede en términos de una variable de métrica monetaria —como gasto o ingreso— cual serie de *Fibonacci*. Dicho de otra forma, es posible que en la ENIGH tengamos sólo a la parte baja o inferior de la clase alta, constituida en buena medida por élites de asalariados más no conglomerados definidos por aquellos hogares cuya percepción central de ingresos sea la renta de la propiedad, es decir —para apelar a la vieja terminología— dueños de medios de producción.

Esta omisión es relevante en términos de la proporción que estas observaciones, no integrables a la muestra, tienen de la riqueza nacional, el ingreso o el gasto corriente, pero ciertamente son un efecto marginal en términos de hogares y personas a la hora de cuantificar de manera demográfica a las clases sociales, que es lo que nos ha ocupado aquí.

13 Lo que no se toma en cuenta entre quienes recomiendan resolver el problema del truncamiento con la asignación de una probabilidad 1 es que el factor de expansión de las encuestas probabilísticas (es decir, el coeficiente que multiplica cada observación —hogar— considerando a cuántos más similares representa fuera de la muestra) es el recíproco de la probabilidad de aparición en muestra. De modo que al multiplicarse por 1 (esto es, 1/1), ello obligaría a tener un censo permanente de ricos como complemento de la ENIGH, pues ninguno de esos hogares podría representar a otro más que a sí mismo. La alternativa de tenerlos con certeza, pero multiplicarlos por el inverso de su bajísima probabilidad original en muestra es inaceptable porque implica multiplicar cada hogar de esos por $1/(x \longrightarrow 0)$, lo que equivale, después de la expansión, a postular la existencia de una magnitud enorme de hogares análogos al de la muestra, lo cual no correspondería a la realidad. En ese sentido, es mejor tomarse en serio esa probabilidad de no aparecer en la muestra y no violar esa condición. (sobre la problemática del truncamiento en las ENIGH ver el trabajo de Fernando Cortés y Rosa María Rubalcava citado por Leyva Parra, 2005).

14 El otro tipo de truncamiento en una encuesta tipo ENIGH se da en el extremo opuesto, es decir, entre los más pobres de los pobres, dado que estos levantamientos cubren viviendas particulares, mientras que hay gente que duerme en refugios, viviendas móviles o también en lugares no construidos para ser habitables. De acuerdo con las cifras del Censo de Población y Vivienda 2010, el monto total de población que pernocta en este tipo de lugares, y por ende no representable por una encuesta tipo ENIGH, ascendería a unos 66.3 mil casos. Esta omisión tampoco deja de ser marginal, pues significa exactamente una décima de punto porcentual del estimado que aquí se ofrece de clase baja y poco más de una décima del monto de pobres que estima el CONEVAL, ver Censo de Población y Vivienda/Tabulados del Cuestionario Básico/Viviendas habitadas y ocupantes por entidad federativa, tipo y clases de vivienda en http://www3.inegi.org.mx/sistemas/tabuladosbasicos/default.aspx?c=27302&s=est

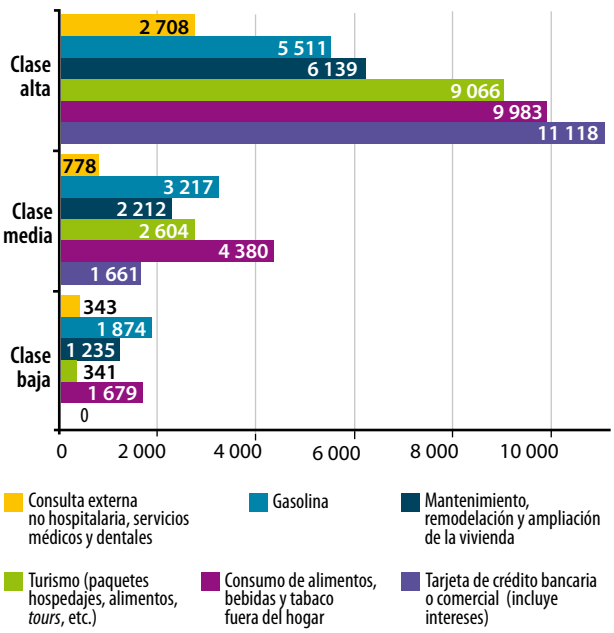
4. Panorámica comparativa de las clases sociales

4.1 Gastos corrientes

A partir de las magnitudes absolutas, que muestra la gráfica 7 en seis rubros de gasto o erogaciones, lo primero que resalta es que la clase baja no tiene acceso a tarjeta de crédito, mientras que lo que la alta abona a este instrumento financiero es una cifra 6.7 superior a la clase media. Los gastos en turismo siguen en cuanto a magnitud de brechas, pues las erogaciones de la alta son 3.5 veces superiores a las de la clase media y 26.5 a las de la baja. Por su parte, la diferencia que hay en gasto de salud (no hospitalarios) de la clase alta en relación con la clase media es, asimismo, de 3.5 veces y 7.9 con respecto a la baja. En el que se refiere a mantenimiento y remodelación de vivienda, la clase alta gasta 2.8 veces más que la media y casi cinco más que la baja. La brecha con respecto a la clase media en consumo de alimentos y bebidas fuera del hogar es menor que la antecedente (2.3), pero con respecto a la clase baja es de 5.9 veces. En cuanto consumo de gasolina, en cambio, las diferencias no son tan grandes: la clase alta desembolsa 1.7 veces más que la media y 2.9 más que la baja (es importante resaltar, por lo demás, que casi una quinta parte de la clase baja tiene acceso a algún tipo de vehículo, ya sea automóvil o una camioneta cerrada con cabina, mientras que, en contraste, en el 2010 no tenía acceso, en ningún caso, ni a internet ni tampoco a computadora u ordenador personal en casa).

Gráfica 7

Gasto promedio trimestral de hogares que ejercieron el gasto en cada concepto



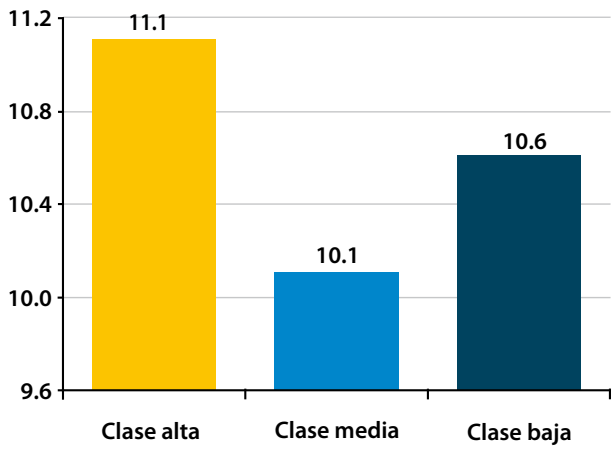
4.2. Mercado de trabajo

El porcentaje de hogares en el 2010 donde ningún integrante percibía ingresos por remuneraciones al trabajo era de 11.1% para la clase alta, 10.1% para la media y 10.6% para la baja (ver gráfica 8). Lo anterior no significa que no haya percepción de ingresos en absoluto o que todos esos hogares fueran de desempleados: muchos de esos hogares pudieron percibir ingresos por otras fuentes (renta de la propiedad) y/o también transferencias (pensiones de divorcio, jubilación, remesas o asimismo transferencias por parte de algún programa gubernamental).

Descontando esos hogares, es decir, centrándonos exclusivamente en aquéllos donde hay una percepción de ingresos por realizar una actividad

Gráfica 8

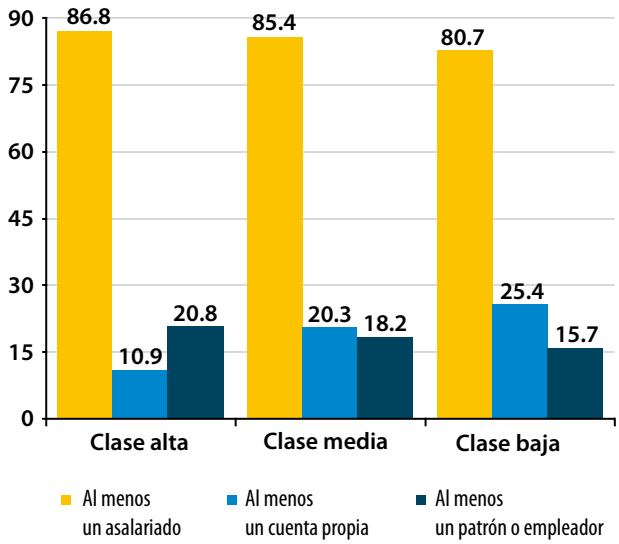
Porcentaje de hogares sin vínculo a las remuneraciones provenientes del trabajo



económica, se tiene que en poco más de una quinta parte de ellos en la clase alta (20.8%) hay un patrón o empleador contra 18.2% en la media y 15.7% en la baja; estos porcentajes descendentes se revierten cuando se mira el porcentaje de hogares con al menos un integrante laborando por su cuenta, es decir, quienes lo hacen sin utilizar empleados remunerados en su proceso de trabajo: para esta última modalidad, mientras se detecta 10.9% de hogares en la clase alta, la proporción asciende a 20.3% en la media y a 25.4% en la baja. Es claro que, a menos de que se trate de profesionistas independientes, quienes de esa manera participan de la actividad económica quedan ligados al ámbito del autoempleo o de los micronegocios, muchos de ellos de carácter tradicional. Trasladando ahora la mirada a la inserción en el mercado de trabajo por la vía asalariada se tienen, otra vez, los porcentajes más altos de hogares con al menos un participante bajo esta modalidad entre la clase alta (86.8%) seguido por la media (85.4%) y luego por la baja (80.7%). Si bien la suma de modalidades detectadas para cada clase rebasa el 100% por el hecho de que pueden darse combinaciones en el seno de hogares con más de un perceptor de ingresos, sin duda es la comparativamente elevada proporción del autoempleo entre la clase baja lo que determina que presente la menor proporción de hogares con un integrante asalariado de las tres clases sociales (ver gráfica 9). La observación recurrente que hace Gabriel Zaid de que en México hay asalariados ricos y emprendedores pobres no deja de tener sustento.[15]

Gráfica 9

Porcentajes de hogares perceptores de ingresos del trabajo en cada clase social según modalidades de inserción económica de sus integrantes

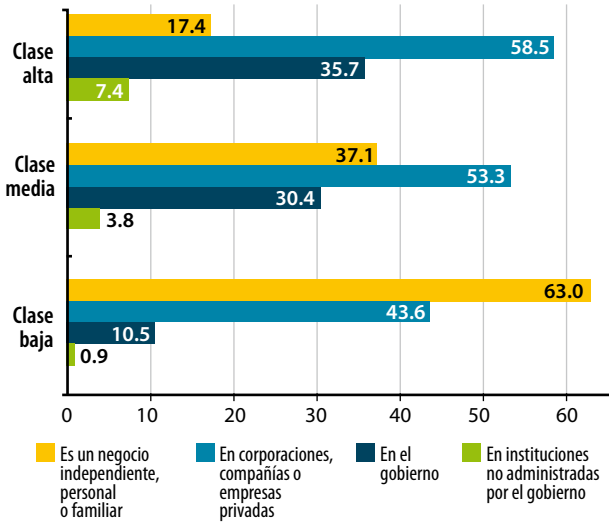


Al aislar ahora dentro de cada clase social a los hogares a cuyo interior por lo menos había una persona integrada al mercado de trabajo asalariado, resalta que, en el caso de la clase baja, 63% laboraba para negocios independientes, personales o familiares; por su parte, en 37.1% de los hogares con asalariados de clase media había alguien que tuviera como fuente de trabajo a este tipo de negocios, mientras que sólo 17.4% en la alta con alguien laborando presentaba ese vínculo. En contraste, 58.5% de los hogares con asalariados de clase alta tenía a alguien trabajando para corporaciones, compañías o empresas privadas constituidas en sociedad, contra 53.3% de la media y 43.6% de la baja. Pero lo que llama la atención es que, entre los hogares de clase alta con al menos una persona integrada al mercado de trabajo, 35.7% contaba con alguien laborando para la administración pública, frente a 30.4% en el caso de la clase media y 10.5% de la baja; las proporciones en lo que se refiere a tener como fuente de trabajo a instituciones autónomas (como universidades, institutos de investigación, poder judicial y otros organismos) eran de 7.4, 3.8 y 0.9%, en ese orden (ver gráfica 10). Cabe observar, de nueva cuenta, que la suma de porcentajes para cada clase supera el 100% por el hecho de que puede haber más de una persona laborando asalariadamente en los hogares, de modo que ello permite combinaciones (por ejemplo, una pareja donde ella labora para la iniciativa privada y él, para el gobierno).

15 Esta observación la ha esgrimido Gabriel Zaid con agudeza desde la brillante colección de ensayos escritos entre finales de la década de los 70 agrupados en ese libro único, provocador e incisivo, *El progreso improductivo*, y en diversos artículos publicados en *Vuelta*, *Contenido*, *Letras Libres* y el *Diario Reforma* a lo largo de muchos años. No está de más señalar que muy pocos han sacado conclusiones al respecto o tomado el toro por los cuernos: un efecto quizás del influjo inconsciente que ejerce todavía el paradigma marxista cuando se piensa en estratificación social.

Gráfica 10

Porcentaje de hogares con asalariados según el tipo de mercado laboral al que éstos se vinculan



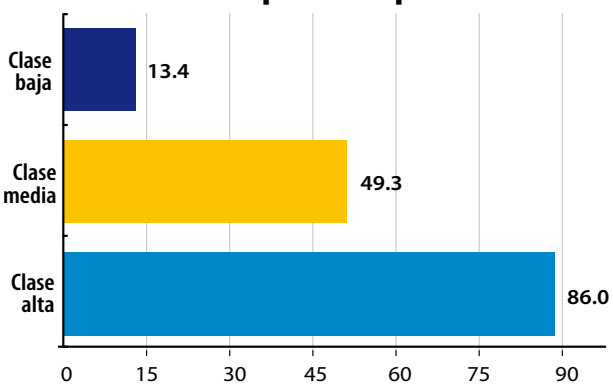
4.3 Nivel de instrucción y formas de acceso a la educación

Tomando como referencia sólo las cabezas de hogares (no su pareja o sus hijos), 86% de la clase alta contaba en el 2010 con educación media superior o superior, contra 49.3% entre quienes encabezan los hogares de clase media y únicamente 13.4% en el caso de jefaturas de hogar de la baja (ver gráfica 11).

Cabe señalar que, entre la clase media urbana, la proporción es de 55.4%, mientras que en la rural, de 29.6%, una brecha considerable que pudiera explicarse, en parte, porque el peso específico del capital humano en los mercados de trabajo en el ámbito rural no sea el mismo que en el urbano. Pero por encima de todo, parece haber un factor sociodemográfico que trasciende a ambos ámbitos siendo propio de la clase media mexicana: una palpable diferencia de escolaridad entre generaciones, misma que ya no es detectable en la alta.

Gráfica 11

Porcentaje de jefes de hogar con educación media superior o superior



Respecto al conjunto de hogares donde hay alguien que asiste a la escuela, en 97.3% de los de clase baja se acude a una institución pública, detectándose, sin embargo, 5.3% con asistencia bajo algún tipo de esquema a escuela privada. La concurrencia a escuela pública sigue siendo preponderante entre los hogares de clase media aunque con un ascenso a 28.7% de hogares con alguien asistiendo a escuela privada (una vez más, pueden darse combinaciones, por ejemplo: un hijo menor asiste a una institución privada y el mayor, a una pública). Sólo en la clase alta la asistencia a una escuela privada es preponderante involucrando a 64.1% de los hogares con alguien en edad escolar al tiempo que 47.8% de los hogares así clasificados tiene un integrante concurriendo a una pública. Todo indica que la escuela pública sigue jugando un papel relevante para la clase media, en particular a nivel medio superior y superior: uno de los pocos espacios ecuménicos socialmente hablando que quedan en el país.

4.4 Algunos aspectos de interés sociológico

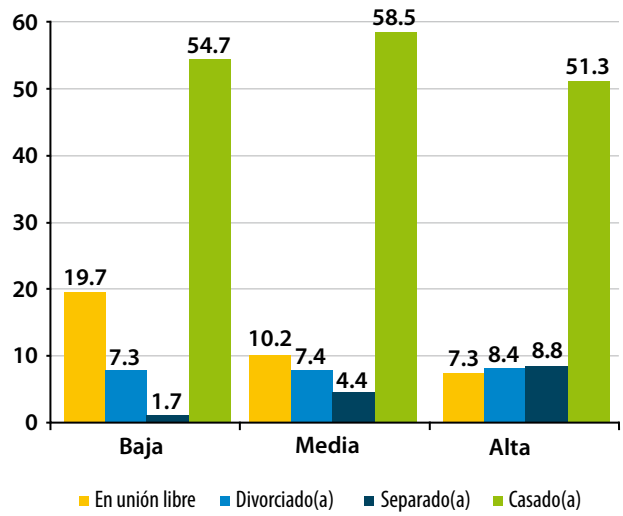
En el ambiente de las relaciones de pareja se percibe una mayor formalidad en los hogares de clase media, donde 58.5% de quienes los encabezan son casados, mientras 10.2% vive en unión libre. En contraste, en la clase baja, las proporciones están en 54.7 y 19.7%, respectivamente. En la alta, el porcentaje de jefes casados es el menor (51.3%), en parte porque el divorcio y la separación tienen una incidencia mayor que en las otras dos clases, no así la unión libre. Probablemente, entre la clase alta, quienes llegan al matrimonio vean a éste con ojos menos conservadores que en la media y más como

un contrato voluntario entre las partes, ¿síntoma de una mayor secularización? (ver gráfica 12)

Una circunstancia en la que poco se repara, pero que tiene un marcado carácter de clase es la discapacidad: en cerca de una quinta parte de los hogares de la baja (19.4%) hay un discapacitado contra 13.2% de los de la clase media y 7.2% de la alta. El que se aproxime a una quinta parte de los hogares en clase baja puede reflejar, en parte, la preponderancia del trabajo físico en dicho segmento social; ya sea que ello genere lesiones, enfermedades o accidentes laborales en el presente o secuelas a futuro, hay que tomar en cuenta en este fenómeno el efecto de un costo asimétrico de la división social del trabajo (ver gráfica 13).

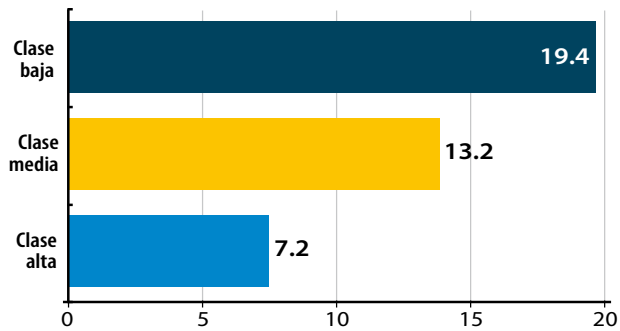
Gráfica 12

Porcentajes según el estado civil del jefe(a)



Gráfica 13

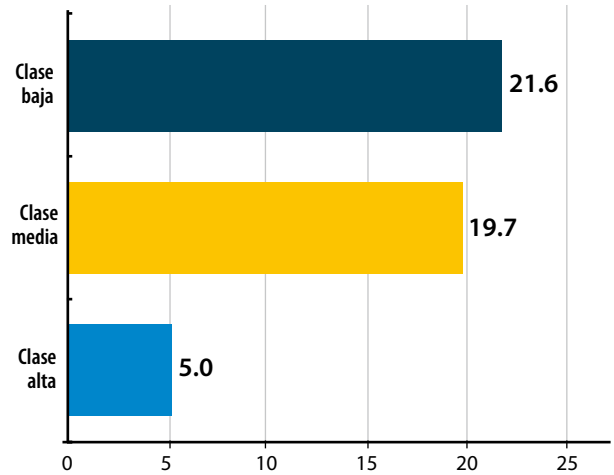
Porcentaje de hogares con al menos una persona que padece discapacidad



Una vez aislando a los hogares con discapacitados, 21.6% de los casos son discapacidades de nacimiento entre la clase baja; 19.7%, en la clase media; contra 5%, en la alta. Más allá de las condiciones de salud de los progenitores que pueden influir en este diferencial de proporciones, también es posible que en ello incidan otros factores, como: los cuidados prenatales, el contar o no con la detección durante la gestación de problemas en el producto, hasta la práctica o posibilidad misma de remitir a instituciones fuera del hogar a cierto tipo de discapacitados (ver gráfica 14).

Gráfica 14

Porcentaje de hogares con personas con discapacidad de nacimiento en el total de hogares con personas con discapacidad

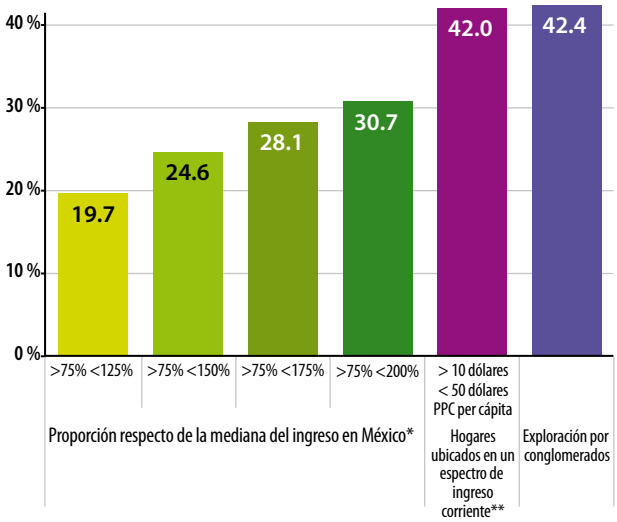


4.5 Comparativo de resultados con otras mediciones realizadas para México

Antes de concluir, vale la pena tener presente las magnitudes a las que se llega en cuanto a la proporción de hogares de clase media en México con distintas metodologías que guardan en común una misma fuente de referencia: la ENIGH. En la gráfica 15, las barras en verde muestran distintos porcentajes en el total de hogares conforme se va ampliando el rango alrededor de la mediana

Gráfica 15

Hogares de clase media según distintas aproximaciones de medición



* Pressman, Steven. "La Clase Media en Países Latinomericanos". en: Revista Problemas del Desarrollo. 164 (42), enero-marzo de 2011.

** López-Calva, Luis F. y Eduardo Ortiz-Juárez. A Vulnerability Approach to the Definition for the Middle Class. The Word Bank Latin America and the Caribbean Region, Poverty, Equity, and Gender Unit, December 2011.

del ingreso, en concreto, el límite superior del mismo. Al hacerlo se tienen resultados que van de 19.7% de los hogares a 30.7% dependiendo del rango de elección. La barra púrpura, por su parte, y que involucra a 42% de los hogares, es lo que resulta del estudio ya mencionado de considerar que la clase media es la parte del espectro social que presenta una probabilidad inferior a 10% de incurrir en pobreza (López-Calva y Ortiz-Juárez, 2011). Por último, la azul ilustra la magnitud de la presente investigación (42.4% de los hogares). Si bien hay que señalar que el ejercicio de López-Calva y Ortiz-Juárez se fundamenta en la ENIGH 2008 y el que aquí se muestra, en la ENIGH 2010, la coincidencia de resultados con la metodología de vulnerabilidad a la pobreza no deja de ser notable considerando que el enfoque y las estrategias metodológicas son enteramente distintos en una y otra aproximación. Nada garantiza que esta convergencia siga manteniéndose en ejercicios futuros, pero sí cabe subrayar que son, hasta ahora, dos metodologías que procuran hacer algo más con la información que simplemente ponerle cotas o fronteras a una variable y luego proclamar como hallazgo lo que se preestablece.

5. Limitaciones del presente estudio y otras vías de exploración

Cuando se aborda lo que podría llamarse una cartografía social con énfasis en la dimensión o espacio que ocupa la clase media, un criterio posible de mejora sería uno que minimizara, hasta donde fuera posible, las decisiones adoptadas por el investigador o el analista; esto es, concebir un método puramente algorítmico de principio a fin. Es difícil decir hasta qué punto sea saludable tal apuesta donde el rol del investigador se reduciría básicamente a correr un programa; por lo pronto, en esta exploración, el agrupamiento final de conglomerados en clases sociales es una decisión que se toma con base en sus características y en su propensión de alejarse o acercarse entre sí, de modo que un método que simplifique ese análisis sería un avance. También, se observará que, puesto que no todo se resuelve en un solo algoritmo, se procede por fases. Otros métodos con la misma filosofía aquí adoptada, es decir, basados en máxima verosimilitud como formas de clasificación con mezclas de distribuciones (Muthén, 2001a y 2001b) podrían ser considerados en estas exploraciones, pero la clave, con el fin de suprimir etapas sucesivas, es que puedan intervenir de manera simultánea variables cuantitativas y categóricas para establecer una clasificación inicial de hogares por grupos de afinidad. Sin duda, ésa fue una dificultad particular que se enfrentó en la fase de conglomeración, donde sólo quedaron involucradas variables cuantitativas.

Otro aspecto a mejorar dado el tema de esta investigación es que se resolviera el ordenamiento de conglomerados por un solo algoritmo sin necesidad de que interviniera un segundo (*Dalenius-Hodges*) tal como se procedió. Un ordenamiento multivariado ciertamente es un reto mayor, pero de encontrársele solución sería, asimismo, decisivo para no segmentar en fases sucesivas y así no salir del algoritmo inicial. Avanzar en esa dirección brindaría un método más puro o, si se quiere, menos híbrido que el aquí presentado.

6. Conclusión

El presente estudio no pretende tener una última palabra en lo que se refiere a la medición de la clase media en México. Lo que plantea es una filosofía de aproximación distinta que reconoce la falta de consensos en cuanto a definiciones o el riesgo de que, aun cuando a ello se llegue, manifieste un carácter demasiado apriorístico o hermético frente a la realidad: hay que dejar, también, que ésta nos hable y nos sorprenda por medio de procedimientos que sean capaces de detectar la configuración de afinidades de las observaciones en estudio —en este caso los hogares— para luego entender qué hay detrás de ellas y no al revés: presuponer que se entiende lo que aún no se analiza. Esto no quiere decir que podamos prescindir de conceptos explícitos o implícitos, pero sí dejar abierta su interacción con los datos. Si de entrada vamos a postular que la clase media supone ciertos niveles de ingresos o gastos, de escolaridad o de jerarquías ocupacionales clausuramos toda posibilidad de captar que pueden establecerse diferencias entre los hogares en México a partir de otros elementos y no sólo éstos.

La exploración que se ha emprendido deja abierta dicha posibilidad, esto es, que nuestra comprensión de los fenómenos no termine ahí hasta donde llegaron nuestras definiciones de punto de partida. Creemos que ésta es en esencia la diferencia entre una filosofía bayesiana y una platónica-aristotélica y nos atrevemos a decir que la segunda no deja aún de proyectar su influencia en las ciencias sociales, en particular en lo que concierne al abordaje de la temática que aquí nos ocupa.

Este ejercicio tampoco pretende explicar, sólo describe. Es abierto porque no confunde lo segundo con lo primero. Comprender la heterogeneidad social —al menos en México— comienza a antojarse como un reto mucho mayor de lo que parece. Es por ello que aquí no se procuró ignorar o prejuzgar la estructura subyacente a los datos bajo estudio y se adoptó una metodología para tomarla en cuenta. El resultado al que se llega nos indica que aún es prematuro proclamar que México —al término de la primera década de este siglo— sea un país mayoritariamente de clase media. El dato puntual no será el único posible, pero tampoco creemos que esta conclusión cambie. En el viaje a Ítaca acaso la travesía vale más la pena que el destino final así que, por lo pronto, ya es algo tener una narrativa de lo que se puede hacer y refinar en exploraciones futuras.

Fuentes

Angus, Deato. *Franco Modigliani and the Life Cycle Theory of Consumption* (PDF). Research Program in Developing Studies and Center for Health and Weel Being. Princeton University, March 2005. Retrived 2014-08-09.

Atkinson, Anthony and Andrea Brandolini. "On the identification of the Middle Class", en: *ECINEQ working paper series*. WP2011-217. September 2011.

Banerjee, Abhijit and Esther Duflo. "What is Middle Class About the Middle Classes Around the World?", en: *MIT, Department of Economics Faculty Research Paper*. MA, USA, Cambridge, December 2007.

Birdsall N., C. Graham, and S. Pettinato. "Stuck in a Tunnel: Is Globalization Muddling the Middle?", en: *Brookings Institution Center Working Paper* No.14. Washington, D.C., 2000.

Bordieu, Pierre. *Poder, derecho y clases sociales*. 2ª. edición. Bilbao, Descleé de Brower, 2001.

_____ *La distinción*. Primera edición. Madrid, Taurus, 1984.

Bussolo M., R. De Hoyos, and D. Medvedev. "The Future of Global Income Inequality", en: Estache, A. and D. Leipziger (eds.). *Stuck in the Middle*: *Is Fiscal Policy Failing the Middle Class?* Washington D.C., Brookings Institution Press, 2009.

Chun, Natalie. "Middle Class Size in The Past Present and Future: a description of trends in Asia", en: *ADB Economic Working Series*. No. 217. September 2010.

CONEVAL. *Medición de la pobreza*: *resultados a nivel nacional y por entidad federativa 2010-2012*. Cuadro resumen, julio del 2012. http://www.coneval.gob.mx/Medicion/Paginas/Medici%C3%B3n/Pobreza%202012/Pobreza-2012.aspx

De la Calle, Luis y Luis Rubio. *Clasemediero*: *pobre no más, desarrollado aún no*. México, DF, Centro de Investigación para el Desarrollo, AC, 2010.

Easterly, W. "Middle Class Consensus and Economic Development", en: *Journal of Economic Growth*. 6 (4), 2001.

Ferreira, Francisco H. G., Julián Messina, Jamele Rigolini, Luis F. López-Calva, María Ana Lugo y Renos Vakis. *La movilidad económica y el crecimiento de la clase media en América Latina. Estudios sobre América Latina y el Caribe*. Washington, Banco Mundial, 2013.

Friedman, Milton (1956). *"A Theory of the Consumption Function"* (PDF). Princeton, NJ: Princeton University Press. Retrived 2014- 08-09.

Goldthrope, John H., and Abigail McKnight. "The Economic Basis of Social Class", en: Morgan, Stephen L., David b. Grusky and S. Gary (eds.). *Fields Mobility and Inequality: Frontiers of Research from Sociology and Economics*. Standford, CA, University Press, 2006.

INEGI. *Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) 2010.* México, INEGI.

_____ *Censo de Población y Vivienda 2010. Tabulados básicos*. México, INEGI.

Kharas, H., and G. Gertz. *The New Global Middle Class: A Cross-Over from West to East*. Washington, D.C., Brookings Institution Press, 2010.

Leyva-Parra, G. "El ajuste del ingreso de la ENIGH con la contabilidad nacional y la medición de la pobreza en México", en: Székely, Miguel (coord.). *Números que mueven al mundo: la medición de la pobreza en México*. , México, DF., SEDESOL-CIDE-ANUIES, Miguel Ángel Porrúa, 2005.

López-Calva, Luis F., Jamele Rigolini, and Florencia Torche. "Is There Such Thing As Middle Class Values: class differences, values and political orientations in Latin America", en: *Policy Research Paper*. No. 5874. Washington D.C., The World Bank Poverty, Equity and Gender Unit, November 2011.

López-Calva, Luis F. y Eduardo Ortiz-Juárez. "A Vulnerability Approach to the Definition of the Middle Class", en: *Policy Research Working Paper*. No. 5902. Washington D.C., The World Bank Poverty, Equity and Gender Unit, December 2011.

Metzgar, Jack. "Are the Poor Part of the Working Class or in a Class by Themselves?", en: *Labour Studies Journal*. Vol. 35, Num. 3, September 2010.

Muthén, B. "Latent variable mixture modeling", en: Marcoulides, G. A. y R. E. Schumacker (eds.). *Structural Equation Modeling*. Mahwah, N.J.: Lawrence Erlbaum Associates, 2001a, pp. 1-33

_____ "Second-Generation structural equation modeling with a combination of categorical latent variables: New opportunities for latent class/latent growth modeling", en: Collins, L. M. y A. Sayer (eds.). *New Methods for teh Analysis of Change*. Washington, D.C., 2001b, pp. 289-332.

Popper, Karl R. "El universo abierto: un argumento a favor del indeterminismo", en: *Tecnos*. Madrid, 1984.

_____ *Conjeturas y refutaciones (1962)*. 2.ª edición. Madrid, Paidós, 1981.

Pressman, Steven. "La clase media en países latinoamericanos", en: *Revista Problemas del Desarrollo*. 164 (42), enero-marzo de 2011.

Ravallion, M., S. Chen, and P. Sangraula. "Dollar a Day Revisited", en: *World Bank Policy Research Working Paper*. No. 4620. Washington, D.C., 2008.

U.S. Department of Commerce. *Middle Class in America*. Report prepared for the Office of the Vice President of the United States Middle Class Task Force, January, 2010.

Weber, Max. *Ensayos de Sociología Contemporánea*. Barcelona, Ediciones Martínez Roca, 1977.

Wittgenstein, Ludwig (1953). *Philosophical Investigations*. Blackwell Publishing, Fourth Edition, U.K. 2009

Zuckerman, Leo. "Pueblo, clientela y ciudadanía", en: *Nexos*. Núm. 389, mayo de 2010.

Anexo

1a. Análisis exploratorio

Es muy recomendable hacer un análisis exploratorio de los datos antes de su procesamiento, pues no sólo ayuda a identificar valores atípicos (*outliers*) o influyentes que puedan sesgar los resultados, sino también descubre correlaciones entre las variables de estudio y con ello, buscar la posibilidad de reducir la dimensión de los datos con técnicas como la de *componentes principales*.

En el presente estudio se aplicaron varias de éstas que orientaron a tomar decisiones importantes sobre la forma de procesar la información, así como su depuración, con el propósito de tener resultados más confiables.

Se tomaron las siguientes decisiones:

- Usando primero la técnica univariada de representaciones gráficas, como la de *box-plot* e histogramas, se distinguieron cuáles valores estaban muy alejados de la distribución media de los datos, y se eliminaron debido a que pudieran afectar los resultados de agrupamiento al causar un sesgo en las estimaciones. Para encontrar datos atípicos multivariantes, se hicieron proyecciones de los datos y se quitaron aquellas observaciones que aparecían muy alejadas de las demás.
- Se usaron gráficos de dispersión de variables por pares para identificar el tipo de relación entre variables. Estas gráficas orientaron la decisión de no realizar transformación de variables.
- Por medio de matrices de correlaciones, y algunos gráficos, notamos que la relación entre las variables no era muy significativa, hecho que se confirmó aún más cuando se realizó un análisis de componentes principales.
- Se vio la necesidad de estandarizar los datos para homogenizar las unidades en todas las variables.
- Se confirmaron distintos comportamientos en zona urbana y rural ratificando la hipótesis inicial de analizar estos dominios por separado.

2a. Método basado en modelos

Para fines ilustrativos, en la gráfica 1a se representa un conjunto de datos bidimensionales en los que se forman varios grupos que se identifican visualmente. Bajo el principal supuesto del método de modelos, cada grupo proviene de una subpoblación con ciertas características geométricas que son identificadas de acuerdo con alguna densidad que se representa como $f_k(\theta_k)$. La suma de estas funciones se llama mezcla de densidades, y se relaciona con la siguiente expresión:

$$f(\boldsymbol{x_i}|\boldsymbol{\theta}) = \sum_{k=1}^{G} \pi_k f_k(\boldsymbol{x_i}|\boldsymbol{\theta}) \; \forall \; i = 1,2,..,n \qquad (1)$$

El valor π_k es la proporción de elementos en el grupo k y son tal que $0 < \pi_k < 1$ y $\sum_{k=1}^{G} \pi_k = 1$. La función $f_k(x_i|\theta)$ es la densidad de la observación *m*-dimensional x_i del *k-ésimo* grupo dado el vector de parámetros θ el cual es desconocido.

Es posible tener distintas distribuciones en la misma mezcla. Por ejemplo, un caso muy trabajado por Scott and Symons (1971) es el que la mezcla sea *gaussiana*. Si se considera este caso, la densidad de la distribución *gaussiana* en cada grupo con media μ_k y varianza Σ_k estaría expresado $f_k(x_i|\mu_k, \Sigma_k)$ y la expresión (1) se reescribiría:

$$f(x_i|\boldsymbol{\theta}) = \sum_{k=1}^{G} \pi_k f_k(x_i|\mu_k, \Sigma_k) \; \forall \; i = 1,2,..,n \qquad (2)$$

Así, el conjunto completo de estimaciones es $\theta\left\{\pi_1, \cdots, \pi_G, \mu_1, \cdots, \mu_G, \Sigma_1, \cdots, \Sigma_G\right\}$.

El objetivo general es estimar los parámetros de las distribuciones de la mezcla y clasificar después, las observaciones por sus probabilidades de pertenencia a las distintas poblaciones o grupos según la distribución conjunta dada por (1).

Estimación bayesiana

Para encontrar la estimación de θ, se recurre a la estimación con enfoque bayesiano. El parámetro es considerado como una variable aleatoria, y la inferencia respecto a sus posibles valores se obtiene usando el teorema de *Bayes* obteniendo la distribución de probabilidad del parámetro condicionada a los datos. De esta manera, se obtiene la media o esperanza de la distribución y, con ello, la estimación puntual del parámetro. La distribución *a posteriori* de θ dada la muestra es:

$$P(\theta|X) = \frac{P(\theta)P(X|\theta)}{P(X)} \propto P(\theta)P(X|\theta) \qquad (3)$$

donde $P(\theta)$ es la distribución inicial, o a *priori*, que puede establecerse como una función no informativa para evitar algún prejuicio del investigador, $P(X|\theta)$ es la distribución conjunta que proporciona probabilidades de valores muestrales, al considerarla como función de θ se convierte en una función de verosimilitud (ver Jasra, A. *et al*., 2005), es decir:

$$P(\theta|X)) \propto \; prior \; x \; verosimilitud \qquad (4)$$

La expresión anterior se lee: "la distribución posterior $P(\theta|X)$ del parámetro dado los datos es proporcional a la información a *priori* de $P(\theta)$ veces la información de los datos"; en otras palabras, dados los valores en la muestra, se buscan los parámetros de la población que más posibilidades tengan de representar a la población que generó a la muestra.

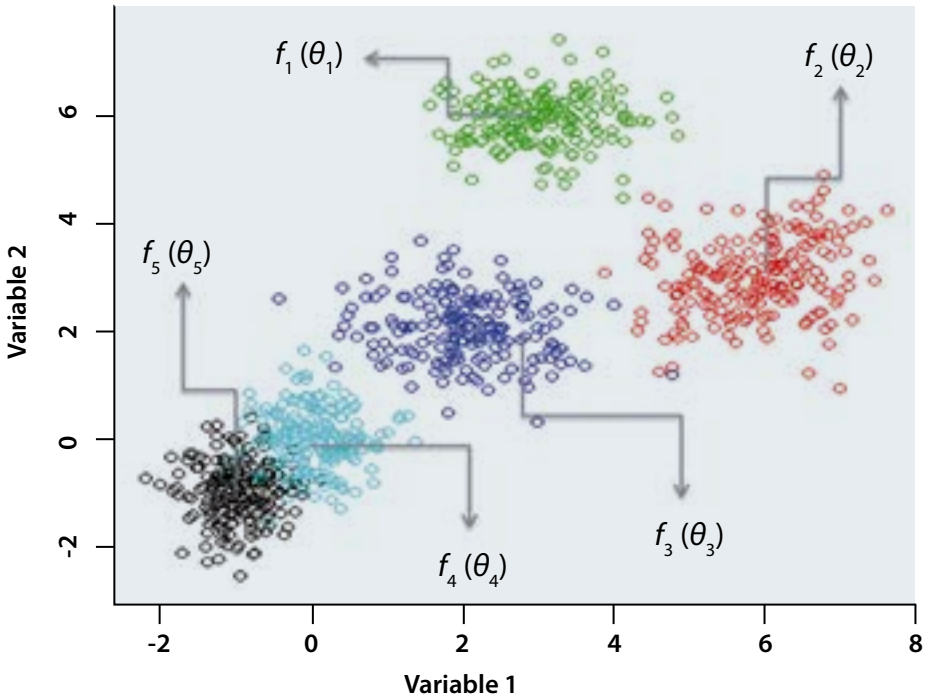
El algoritmo más utilizado para obtener estimaciones de máxima verosimilitud para los parámetros de la mezcla es el *algoritmo EM*, cuyas iniciales provienen de *Expectation-Maximization* (Dempster *et al*., 1977, ver McLachlan y Krishnan,1977). Hay que comentar que existen adecuaciones a este algoritmo para hacerlo más eficiente, (ver Fraley, Chris, 2007). Este algoritmo, junto con sus adecuaciones, se utilizó para este trabajo.

Reparametrización

Una de las grandes ventajas del método basado en modelos es que toma en cuenta la correlación de las variables para realizar las agrupaciones;

Gráfica 1a

Representación de agrupación en un espacio de dimensión dos, cada grupo es representado por un color y una función de distribución que pueden ser distintas



gracias a ello, el método es capaz de identificar grupos con distintas formas, orientación y volumen. De esta manera, cuando se estima θ se toma en cuenta una reparametrización usando la descomposición espectral, propuesta por Banfield y Raftery (1993) de la matriz de covarianza en cada subpoblación o en cada grupo, quedando de la siguiente forma:

$$\Sigma_k = \lambda_k D_k A_k D_k^t \; con \; k = 1,2,..,G \qquad (5)$$

donde:

Σ_k = a la matriz de covarianza de la *k-ésima* subpoblación.

D_k = determina la *orientación* de los elipsoides (grupos), matriz ortogonal de eigenvectores.

A_k = sirve para identificar la *forma* de la distribución, es una matriz diagonal compuesta por los eigenvalores de Σ_k.

λ_k = es un escalar e identifica el volumen.

A partir de lo anterior, es posible hacer una serie de supuestos sobre las matrices en cada grupo; éstas pueden ser iguales o, en caso extremo, todas diferentes. A estas caracterizaciones las llamaremos *plantillas* o *modelos*; por ejemplo, si se hace el supuesto de que las matrices de covarianza son iguales en todos los grupos, entonces la matriz es caracterizada así: $\Sigma_k = \sigma^2 \boldsymbol{I}$ para toda $k = 1, ..., G$, donde $\boldsymbol{I}$ = matriz identidad, y significa que no existe correlación entre las variables de los datos. Esta restricción es la misma que supone el algoritmo que usa la minimización de suma de cuadrados como criterio de paro para el procedimiento de agrupación.

Por otro lado, el caso opuesto a este ejemplo es la caracterización menos parsimoniosa donde se da la libertad a los parámetros de la matriz de covarianzas que varíen y que sean desiguales en todos los grupos.

Entre estos dos ejemplos existirán casos intermedios, resultado de la combinación de la variación de parámetros de la matriz Σ_k; por ejemplo, si $\Sigma_k = \lambda_k DAD^t$, entonces se tendrá un modelo que varía en volumen, pero tienen la misma forma y orientación. Si $\Sigma_k = \lambda D_k A D_k^t$, entonces se tendrá un modelo que varía en forma y orientación, pero con igual volumen o proporción.

La tabla 1a muestra 10 representaciones de distintas estructuras de matriz de covarianza. La representación *EVI* indicaría un modelo donde todos los grupos tienen el mismo volumen (*E*, *equal*); la forma de los grupos puede variar (*V*, *varing*) y la orientación es *I* idéntica (*I*, *Identity*) que corresponde a una distribución diagonal. En total, tendremos 10 formas o 10 distintos modelos posibles de agrupar los datos. La columna que se refiere a la distribución indica los contornos de las densidades; para el caso de una mezcla de normales, se dice que son distribuciones elipsoidales, pero si la matriz de covarianza es restringida a ciertos valores —como en el caso cuando la matriz de covarianza no varía en cada grupo— entonces los contornos son esféricos, o bien, si la matriz es de la forma $\Sigma_k = \lambda_k A$, el contorno está alineado a un eje y la distribución es llamada diagonal.

El criterio *Bayesian Information Criteria*

La decisión sobre cuál plantilla o modelo es el que más se ajusta a la población de estudio se toma bajo el principio de contraste de hipótesis, pero con un enfoque bayesiano. Cada modelo o plantilla se ve como una hipótesis que será contrastada con los demás modelos y se elegirá aquél con máxima probabilidad *a posteriori*. Para encontrar dicha probabilidad, se hacen ciertas aproximaciones de los términos que la componen y se desprende una expresión en términos de *log-verosimilitud* que pondera la desviación del modelo con el número de parámetros, esta expresión es llamada criterio BIC[16] (Schwarz, 1978) y es usado en un amplio número de aplicaciones (e.g. Dasgupta and Raftery, 1998; Fraley and Raftery, 1998 y 2002); en este documento lo expresamos como:

$$BIC(M) = 2 * logovero_{\,M}\left(X, \hat{\theta}^*\right) - (\#parametros)_M * \log(n) \qquad (6)$$

donde $\text{logovero}_{\text{M}}\ (X,\hat{\theta}^*)$ es el valor máximo de la *log verosimilitud* de los datos usando algún modelo de los 10 disponibles; el valor $(\#\ \text{parametros})_M$ es el número de parámetros independientes que son estimados en el modelo M y n, el número de observaciones. Puede demostrarse que el criterio BIC es consistente de manera que la probabilidad de seleccionar el modelo correcto tiende a 1 si crece el tamaño muestral.

Se obtendrá el valor del criterio BIC para cada plantilla y para diferentes números de grupos. Se sugiere elegir el modelo o plantilla que en combinación con el número de grupos maximice[17] el criterio BIC, esto puede apreciarse mejor en la gráfica 2a. Es importante decir que este criterio guía al investigador para tomar una decisión, no necesariamente debe de considerarse como una regla automática.

3a. *Software* disponible

Este trabajo fue desarrollado en el ambiente R y se utilizó una de sus librerías llamada *MCLUST*; en ella se encuentra implementado el método basado en modelos (disponible en http://www.stat.washington.edu/mclust); asimismo, cuenta con adaptaciones de algoritmos conocidos que mejoran las estimaciones, por ejemplo, resuelve situaciones de no-convergencia o de soluciones no factibles que se pueden presentar en el *algoritmo EM*, ver Fraley, Chris y A. E. Raftery (2002).

16 Este criterio no es una probabilidad, es una cantidad que puede variar en distintas escalas.

17 Algunos autores definen el BIC con signo contrario a la expresión (3). En estos casos, el valor más pequeño (más negativo) es el que se toma como referencia para la elección del mejor modelo.

Gráfica 2a

Valores BIC para distintos modelos y distinto número de grupos; se elige la combinación que maximice el criterio

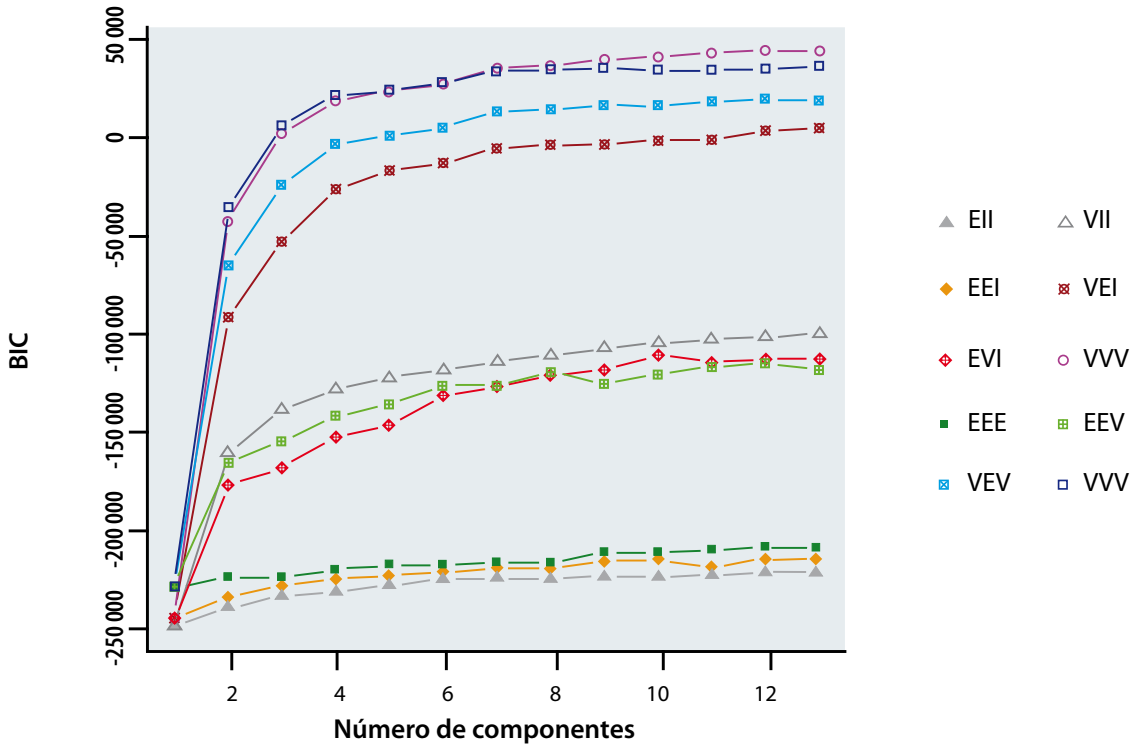


Tabla 1a

Parametrización de la matriz de covarianza y su relación con la forma geométrica

Identifier	*Model*	*Distribution*	*Volume*	*Shape*	*Orientation*
EII	λI	Esférica	*Equal*	*Equal*	NA
VII	$\lambda_k I$	Esférica	Variable	*Equal*	NA
EEI	λA	Diagonal	*Equal*	*Equal*	*Coordinate axes*
VEI	$\lambda_k A$	Diagonal	Variable	*Equal*	*Coordinate axes*
EVI	λA_k	Diagonal	*Equal*	Variable	*Coordinate axes*
VVI	$\lambda_k A_k$	Diagonal	Variable	Variable	*Coordinate axes*
EEE	λDAD^t	Elipsoidal	*Equal*	*Equal*	*Equal*
EEV	$\lambda D_k A D_k^t$	Elipsoidal	*Equal*	*Equal*	Variable
VEV	$\lambda_k D_k A D_k^t$	Elipsoidal	Variable	*Equal*	Variable
VVV	$\lambda_k D_k A_k D_k^t$	Elipsoidal	Variable	Variable	Variable

4a. Conclusiones generales

Como un resumen del método de agrupación basado en modelos, podemos destacar lo siguiente:

- Se hace la suposición de que los grupos que pudieran formarse provienen de una subpoblación con ciertas características geométricas que se identifican de acuerdo con alguna distribución basada en su densidad.
- Toma en cuenta la correlación de las variables para realizar las agrupaciones. Gracias a ello, puede identificar grupos con distintas formas, orientaciones y tamaños.
- El método añade un criterio que sugiere, con base en sus probabilidades *a posteriori*, la mejor característica geométrica (forma, tamaño y orientación) que se adapte a los datos.
- El problema de determinar el número de grupos se resuelve simultáneamente eligiendo, también, el mejor modelo o plantilla que se adapte a los datos según el criterio BIC.

Además, por la naturaleza del método, podemos conocer las incertidumbres de los elementos clasificados en algún grupo, es decir, es posible saber qué probabilidad tiene el elemento de ser asignado a cualquier *cluster* y estimar el error de medición. Otra ventaja es que el método también cuenta con la capacidad de identificar la presencia de ruido y *outliers*, suponiendo que es un grupo distinto a los demás, modelado con una distribución *Poisson*. Para más detalles, consultar las fuentes mostradas.

Fuentes

Dasgupta, A. and A. E. Raftery. "Detecting features in spatial point processes with clutter via model based clustering", en : *Journal of the American Statistical Association*, 93:294-302,1998.

Banfield, J. D. y A. E. Raftery. "Model-based Gaussian and Non-Gaussian Clustering", en: *Biometrics*. 49, 1993, pp. 803-821.

Damaris, Pascual. *Algoritmos de agrupación basados en densidad y validación de clusters*. Tesis doctoral. Castellón, Marzo del 2010.

Dempster, A. P., N. M. Laird and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm" (with discussion), en: *Journal of the Royal Statistical Society*. Series B (Methodological), Vol. 39, No.1 ,1977, pp.1-38.

Fraley, Chris. "Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering", en: *Journal of Classification*. 24, 2007, pp.155-181.

Fraley, Chris. y A. E. Raftery. "How Many Clusters? Which Clustering Method? – Answers via Model-based Cluster Analysis", en: *Computer Journal*. 41, 1998, pp. 578-588.

_____"Mclust: Software for Model-based Cluster Analysis", en: *Journal of Classification*. 16, 1999, pp. 297-306.

_____ "Model-based Clustering, Discriminant Analysis and Density Estimation", en: *Journal of the American Statistical Association*. 97, 2002, pp. 611-631.

_____ "Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering", en: *Journal of Classification*. 24, 2007, pp. 155-181.

Jasra, A., C. C. Holmes y D. A. Stephens. "Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modelling", en: *Statistical Science*. 20, 2005, pp. 50-67.

McLachlan, G. J. and T. Krishnan. T*he EM Algorithm and Extensions*. NewYork. Wiley, 1997.

Peña, D. *Análisis de datos multivariantes*. McGraw-Hill Interamericana, 2004.

Schwarz, G. "Estimating the dimension of a model", en: *Ann. Statist.* ,6, 1978, 461-464.