La integración de los microdatos censales de América Latina: el proyecto IPUMS-América Latina

Robert McCaa, Albert Esteve, Steven Ruggles y Matthew Sobek*

Gracias al pionero esfuerzo del doctor Gustavo Cabrera y de otros grandes próceres de la demografía, en América Latina sobrevive un vasto archivo de microdatos censales; sin embargo la mayor parte de ellos se mantiene inaccesible a los investigadores.

En la trayectoria académica y científica del profesor Cabrera ha sido constante su preocupación por las fuentes de información. Hoy el proyecto Integrated Public Use of Microdata Series para América Latina (IPUMS-AL) acomete con ímpetu la difícil tarea de integrar los microdatos censales de esta región, haciendo uso intensivo y extensivo de las nuevas tecnologías pero, sobre todo, contagiándose del empeño que instituciones y eminentes demógrafos latinoamericanos han dedicado a mejorar la calidad y a preservar estos datos, que constituyen sin lugar a dudas los tesoros estadísticos de América Latina.

El proyecto IPUMS-América Latina cuenta con el soporte económico necesario para integrar esos microdatos en una única base de datos armonizada que estaría destinada a la investigación académica y a la que se podría acceder desde Internet. Los microdatos censales de 1960, 1970, 1990 y 2000 de México ya han sido integrados (www.ipums.org/international) como resultado del trabajo colectivo desarrollado en el INEGI, socio fundador del proyecto, de destacados demógrafos mexicanos, y del Population Center de la Universidad de Minnesota.

Palabras clave: censos, microdatos, América Latina.
Fecha de recepción: 7 de noviembre de 2003.
Fecha de aceptación: 16 de diciembre de 2004.

Integrating Census Microdata on Latin America: the IPUMS-Latin American Project

Thanks to the pioneering efforts of Dr. Gustavo Cabrera and other leading demographers, Latin America contains a vast archive of census microdata, the majority of which, however, are inaccessible to researchers.

Throughout his academic and scientific career, Professor Cabrera was constantly concerned with information sources. The Integrated Public Use of Microdata Series for Latin America (IPUMS-AL) has embarked on the difficult task of integrating the census microdata from this region by making intensive and extensive use of new technologies, but above all, by infusing them with the determination with which Latin American

* Minnesota Population Center. Correo electrónico: rmccaa@tc.umn.edu.

institutions and eminent demographers have sought to improve the quality and ensure the preservation of these data, which undoubtedly constitute one of Latin America's statistical treasures.

The IPUMS-Latin American project has the financial basis required to incorporate these microdata into a single data base that will be used for academic research and be accessible via the Internet. The census microdata of 1960, 1970, 1990 and 2000 on Mexico have already been integrated (www.ipums.org/international) as a result of the collective work undertaken by INEGI, a founding member of the project, leading Mexican demographers, and the University of Minnesota Population Center.

Key words: censuses, microdata, Latin America.

Los tesoros estadísticos de América Latina

Los microdatos censales son un recurso de gran valor para la investigación en ciencias sociales (McCaa y Ruggles, 2002) por un doble motivo: por su condición de microdatos, registros individuales que permiten explorar simultáneamente las características de los individuos, familias, hogares y viviendas en que residen, y porque proceden del censo, fuente estadística sin parangón, pues ninguna otra ofrece una densidad muestral, profundidad cronológica, y cobertura geográfica comparables. No en vano el censo conserva la mayor representatividad a escala nacional.

En el caso preciso de México, los esfuerzos del profesor Cabrera y otros demógrafos han hecho posible que los investigadores mexicanos se aprovechen de las muestras censales registradas desde la década de los sesenta. Primero fue una muestra de 1.5% de individuos del Censo de 1960. Después, una muestra de 1% del Censo de 1970. En cuanto al Censo de 1980, no se pudo levantar a tiempo una muestra a causa de la destrucción de los registros originales a raíz del temblor de 1985. Para el Censo de 1990, el Instituto Nacional de Estadística, Geografía e Informática (INEGI) difundió una excelente muestra de 1%, y para 2000 una muestra incomparable de 10%, que estuvo disponible en tiempo récord –poco más de un año después de la conclusión del empadronamiento de casi 100 millones de personas–. Durante todo este periodo, el Consejo Nacional de Población ha custodiado y facilitado el acceso a las muestras mexicanas gracias a la precursora labor que desempeñó el profesor Cabrera mientras presidió esta institución.

América Latina en su conjunto posee la más sustanciosa colección de microdatos censales del mundo, con series completas para las

últimas cuatro décadas. Con más de cien millones de registros, esos microdatos censales ofrecen mayor densidad muestral y alcance temporal que cualquier otro tipo de datos. Además, al compararla con la de otras regiones del mundo, la calidad de estos datos es excepcional. Los censos de América Latina presentan una gran uniformidad pues la región comparte una cultura estadística común, alimentada por cinco décadas de coordinación metodológica, gracias al tesón de dos instituciones: el Comité de Censos de las Américas del Instituto Interamericano de Estadística (Cota) y el Centro Latino Americano y Caribeño de Demografía de la Naciones Unidas (Celade). Cota comenzó su trabajo en la ronda censal de 1950 y continuó con un vigoroso programa de conferencias y seminarios a lo largo de la década de los sesenta. Desarrolló los lineamientos para diseñar los censos que posteriormente fueron adoptados internacionalmente en las rondas censales de 1970 y 1980. Con la excepción notable de Brasil, el resto de los países adoptaron los estándares propuestos. De esta manera se alcanzó el objetivo perseguido: mantener y mejorar la comparabilidad de la información censal en el tiempo y el espacio. Desde sus programas de docencia (dentro de los cuales figura el profesor Cabrera entre la primera cohorte de graduados), talleres, seminarios de entrenamiento y capacitación, publicaciones, y conferencias, Celade ha contribuido significativamente a mejorar la calidad de los censos en América Latina. Para la ronda censal de 1990 Celade tomó el relevo de Cota en la elaboración de lineamientos, y para ello sugirió un diseño estándar de cuestionario que incluyó un grupo de preguntas comunes y una forma homogénea de presentación de las preguntas censales (Celade, 1989).

En 1959 Celade emprendió un ambicioso proyecto orientado a crear un archivo de microdatos censales para toda América Latina y el Caribe (McCaa y Jaspers, 2000). Este proyecto, llamado Operación de Muestras de Censos (Omuece), tenía como objetivos no sólo recolectar y preservar los microdatos y documentación censales, sino también estandarizar una selección de variables para 29 censos levantados entre 1960 y 1976. Pese a las restricciones económicas que lo obligaron a abandonar este proyecto a principios de los años ochenta, Celade continuó recolectando microdatos censales para las naciones de América Latina.[1] Gracias a su perseverancia y empeño, esta

[1] Desde mediados de la década de los ochenta, Celade se comprometió a colaborar en un proyecto de desarrollo de programas de cómputo para proveer tabulaciones de microdatos a pequeñas áreas, cuyo nombre era Recuperación de Datos para Áreas

institución cuenta hoy con la mayor colección de datos censales del mundo. Sin embargo los investigadores han hecho poco uso de estos microdatos debido a que Celade tuvo que restringir su uso a sus oficinas centrales en Santiago de Chile. Para trabajar fuera de este lugar, los potenciales usuarios deben obtener permisos individuales de cada instituto nacional de estadística antes de acceder a los datos.

El Proyecto IPUMS-América Latina

El proyecto IPUMS América Latina (IPUMS-AL) nació con una vocación clara: poner a disposición de la comunidad científica los microdatos censales de América Latina con base en la filosofía de IPUMS. A continuación, antes de tratar los detalles específicos de IPUMS-AL, consideramos oportuno presentar en forma breve sus antecedentes inmediatos.

Los antecedentes de un proyecto

En la mayoría de países los microdatos no están al alcance de los investigadores o su acceso es restringido, razones que explican su escasa utilización. Estados Unidos y Canadá son una excepción al respecto, ya que sus microdatos han estado disponibles desde la década de los sesenta y hoy día son un componente indispensable de la infraestructura en ciencias sociales.

Las Series de Microdatos Censales Integrados de Uso Público (Integrated Public Use Microdata Series-USA) son parcialmente responsables del uso extendido de microdatos censales entre los demógrafos y otros investigadores interesados en el estudio de Estados Unidos. IPUMS-USA, proyecto desarrollado por Ruggles *et al.* (1997) en el Population Center de la Universidad de Minnesota, ha puesto a disposición gratuita de la comunidad científica series de microdatos armonizadas desde 1850 hasta 2000, debidamente documentadas, mediante un sistema de fácil acceso a los datos. Desde su puesta en marcha preliminar en 1995, IPUMS ha sido una de las fuentes de datos demográficos más utilizadas en el mundo entero.

Pequeñas por Microcomputador (Redatam). Hoy día es utilizado por planificadores y analistas gubernamentales en toda América Latina. Sin duda alguna esta iniciativa es un excelente complemento de IPUMS-América Latina.

En 1998, por primera vez, se extendió el paradigma de IPUMS a los censos de Colombia, en una experiencia piloto realizada en estrecha colaboración con el Departamento Nacional de Estadística de Colombia (Dane). Sin duda alguna, Col-IPUMS colocó la primera piedra de lo que más tarde sería una exitosa iniciativa de integración de microdatos a escala internacional: IPUMS-Internacional.

En 1999 la agencia estadunidense National Science Foundation financió la propuesta IPUMS-Internacional, y así dio vida a un proyecto que hoy, en su cuarto año, ha integrado con éxito datos de siete países (China, Colombia, Estados Unidos, Francia, Kenia, México y Vietnam), ha inventariado otras muestras de microdatos en formato de cómputo que se han reunido en el mundo para el periodo 1960-2000 (Hall, McCaa, y Thorvaldsen, 2000), y ha preservado muestras de microdatos de más de cien censos. Además de estos logros, es notable la buena acogida que ha merecido este proyecto entre los académicos.

IPUMS-Internacional es consciente del potencial asociado al uso de los microdatos, razón por la cual trabaja activamente para poner a disposición de la comunidad científica series integradas para el máximo número de países posible, en cooperación con los institutos de estadística nacionales, los centros de investigación y los profesionales de la demografía.

Las participaciones de México y Colombia en IPUMS-Internacional fueron decisivas en la gestación del proyecto IPUMS-AL. En el caso de México, por ejemplo, se integraron datos de cuatro censos, de cuyas características se informa en el cuadro 1. Para la conformación de los datos mexicanos se contó con la ayuda de acreditados demógrafos de este país, cuya asesoría fue esencial para una mejor comprensión de las definiciones y los conceptos censales. Sin duda alguna, la colaboración de los expertos mexicanos fue clave para el éxito de IPUMS-Internacional. Los buenos precedentes de México y Colombia estimularon al Population Center de la Universidad de Minnesota para emprender una nueva iniciativa que integrara los microdatos censales de América Latina.

IPUMS-AL hoy

IPUMS-América Latina es hoy una realidad. En cinco años prevé difundir datos de más de 70 censos de 17 países. Gracias al esfuerzo con-

CUADRO 1
Características de las muestras de los censos de México

Características del censo	*1960*	*1970*	*1990*	*2000*
Título	VIII Censo General de Población y Vivienda, 1960	IX Censo General de Población y Vivienda, 1970	XI Censo General de Población y Vivienda, 1990	XII Censo General de Población y Vivienda, 2000
Agencia censal	Dirección General de Estadística, Secretaría de Industria y Comercio	Dirección General de Estadística, Secretaría de Industria y Comercio	Instituto Nacional de Estadística, Geografía e Informática (INEGI)	Instituto Nacional de Estadística, Geografía e Informática (INEGI)
Universo	Residentes en México; nacidos en el extranjero con más de seis meses de residencia en México, excluye personal diplomático	Residentes en México, excluye diplomáticos extranjeros, incluye diplomáticos mexicanos, militares y marineros, y a sus familias, residentes en otros países	Residentes en México, excluye diplomáticos extranjeros, incluye diplomáticos mexicanos, militares y marineros, y a sus familias, residentes en otros países	Residentes en México, incluye diplomáticos mexicanos y sus familias residentes en otros países; residentes extranjeros, no incluye extranjeros en servicios diplomáticos ni a sus familias. El Censo procuró enumerar a vagabundos, sin techo y trabajadores en tránsito. La versión actual de los datos excluye a las personas

				que viven en el extranjero
De jure o *de facto*	*De jure*	*De jure*	*De jure*	*De jure*
Unidad de enumeración	Vivienda ocupada	Vivienda ocupada	Vivienda ocupada	Vivienda ocupada
Día censal	8 de junio, 1960	28 de enero, 1970	12 de marzo, 1990	14 de febrero, 2000
Periodo de enumeración	8 de junio, 1960	La mayor parte del trabajo fue completado en el día de la enumeración	12 a 16 de marzo, 1990	7 a 18 de febrero, 2000
Cuestionarios censales	Cuestionario separado para cada bloque censal	Cuestionario separado para cada vivienda	Cuestionario separado para cada vivienda	Cuestionario largo y cuestionario corto de vivienda
Tipo de enumeración	Enumeración directa	Enumeración directa	Enumeración directa	Enumeración directa
Responde	Cabeza del hogar	Cabeza del hogar	Cabeza del hogar	Personas de 15 o más años que residan en la vivienda y tengan conocimiento de los otros residentes
Subestimación	No se cuenta con estimaciones oficiales			

(continúa)

CUADRO 1
(continuación)

Características delas muestras	*1960*	*1970*	*1990*	*2000*
Fuente de los microdatos	Celade	INEGI. Realizada a partir de una muestra de cuestionarios censales para obtener los resultados preliminares del censo	INEGI. Realizada a partir de 100% de los microdatos	INEGI. Realizada a partir de 100% de los cuestionarios largos
Diseño muestral	Muestra representativa de individuos	Muestra sistemática de cada centésima vivienda a partir de un punto de inicio aleatorio	Muestra sistemática de viviendas privadas, ordenadas geográficamente para mejorar la precisión. Muestras realizadas por estados	Diseño estratificado por *clusters*; estratificado geográficamente por municipalidades y áreas urbanas. Los *clusters* son definidos como áreas de enumeración (AGEB), bloques de viviendas o localidades. Todas las viviendas dentro de un *cluster* se incluyen en la muestra. La fracción muestral depende de la heterogeneidad demográfica de las municipalidades. Esta muestra fue diseñada

				para ofrecer resultados representativos para las localidades con 50 000 habitantes o más
Unidad de la muestra	Individuos	Viviendas	Viviendas	Áreas de enumeración (AGEB/bloque de viviendas/localidades)
Densidad muestral	1.50%	1%	1%	10.60%
Tamaño muestral registros individuales)	502 800	483 405	802 767	10 099 182
Factor de ponderación	Autoponderadas	Autoponderadas	Autoponderadas	Los factores de ponderación calculados por la agencia estadística deben usarse en la mayoría de los análisis
	Factor de expansión = 66.7	Factor de expansión = 100	Factor de expansión = 100	

Unidades identificadas	*1960*	*1970*	*1990*	*2000*
Viviendas	No disponibles en la muestra	Sí	Sí	Sí
Viviendas desocupadas	No	No	No	No

(continúa)

CUADRO 1
(conclusión)

Unidades identificadas	*1960*	*1970*	*1990*	*2000*
Hogares	No disponibles en la muestra	Sí	Sí	Sí
Individuos	Sí	Sí	Sí	Sí
Viviendas colectivas	No identificadas	No identificadas	No incluidas en la muestra de microdatos	No incluidas en la muestra de microdatos
Poblaciones especiales				Los migrantes no residentes no están incluidos en la muestra de microdatos

Definiciones de unidades identificadas	*1960*	*1970*	*1990*	*2000*
Vivienda	Viviendas ocupadas con entrada independiente usada como albergue	Viviendas ocupadas con entrada independiente usada como albergue	Viviendas ocupadas con entrada independiente usada como albergue	Viviendas ocupadas con entrada independiente usada como albergue
Hogares privados	Grupo de personas, emparentadas o no,	Grupo de personas, emparentadas o no,	Grupo de personas, emparentadas o no,	Grupo de personas, emparentadas o no,

	que viven juntas bajo el mismo techo y comparten los gastos de la comida	que viven juntas bajo el mismo techo y comparten los gastos de la comida	que viven juntas bajo el mismo techo y comparten los gastos de la comida	que viven juntas bajo el mismo techo y comparten los gastos de la comida
Viviendas colectivas	No definidas	No definidas	Edificio usado para acoger personas por razones de asistencia, salud, educación, religión, encarcelamiento o servicio	Edificio usado para acoger personas por razones de asistencia salud, educación, religión, encarcelamiento o servicio

FUENTES: Rabell (2000), Eternod y Trejo (2002), INEGI (2000 y 1993), Dirección General de Estadística (1972 y 1962) e IPUMS-Internacional (2000).

junto de Celade, de los institutos de estadística de esta región del mundo, y del Minnesota Population Center, en julio de 2003 el National Institutes of Health financió el proyecto de integración de los datos de 17 países de América Latina. Con más de 100 millones de registros, que abarcan un periodo de cuatro décadas, la nueva base de datos permitirá a los científicos sociales realizar análisis comparativos para un lapso marcado por intensos cambios. Se trata de la iniciativa regional más ambiciosa que se ha llevado a cabo en este ámbito, llamada a influir significativamente en las ciencias sociales y, concretamente, en campos como la planificación, las políticas públicas en materia de salud, el desarrollo económico y las transformaciones demográficas en aspectos como el envejecimiento, la estructura familiar y las migraciones internacionales.

IPUMS-AL no sólo pretende hacer disponibles los datos censales de América Latina, sino también hacerlos útiles. Incluso donde los microdatos pueden ser obtenidos, la realización de estudios comparativos entre países o periodos históricos resulta un auténtico reto debido a las inconsistencias de las bases de datos y a la deficiente documentación de tales inconsistencias. Por esta razón raramente se desarrolla la investigación comparativa internacional basada en muestras censales homologadas. De conformidad con la filosofía de IPUMS-Internacional, IPUMS-AL reducirá las barreras a la investigación internacional al transformar los microdatos censales de distintos países en una base de datos uniforme y homogénea, proveyendo extensa documentación, y poniendo la información al alcance de los investigadores interesados en forma totalmente gratuita.

En relación con la metodología de trabajo, se replicará el procedimiento utilizado para IPUMS-Internacional. La información se procesa en grupos de tres o cuatro países, cuyos microdatos se difunden cuando están completamente integrados. Se trata de un sistema basado en fases que evita la complejidad logística que supone abarcar muchos censos simultáneamente. La secuencia de procesamiento propuesta es como sigue:

1. México, Colombia, Brasil
2. Costa Rica, Panamá, Chile
3. Argentina, Paraguay, Perú
4. Ecuador, Venezuela
5. República Dominicana, El Salvador, Guatemala
6. Honduras, Nicaragua y eventualmente, si se firman los acuer-

dos generales que gobiernan el uso de los microdatos, Bolivia, Cuba y Uruguay.

En el primer grupo se incluyen 13 censos de tres países, para cuya integración se ha recibido financiamiento con anterioridad, razón por la cual México y Colombia ya están prácticamente integrados y Brasil se encuentra en una fase muy avanzada del proceso. La distribución de los datos de Brasil está programada para principios de 2004. Por lo que al resto de países se refiere, los datos no se distribuirán hasta completar todas las fases de la integración.

Descripción de la base de datos

En el cuadro 2 se presentan los censos que van a ser incorporados en la base de datos. Se incluyen Bolivia y Uruguay, los dos países que aún no han firmado el acuerdo, y Puerto Rico, estado asociado a Estados Unidos y cuyos datos ya son de dominio público. En la parte izquierda del cuadro se informa qué porcentaje de casos está disponible para cada censo. Sin contar la ronda de 2000, en 27 de los 66 censos se conserva 100% de los microdatos. En el resto de los casos la densidad de las muestras oscila entre 1 y 25%. La mayoría de las muestras incompletas pertenece a las rondas censales más antiguas, las de 1960 y 1970. Para la ronda de 1960, tan sólo algunas muestras se preservan en formato electrónico de cómputo, y la mayoría de ellas fueron referidas a los individuos, lo que no permite conocer la composición y estructura de los hogares. Aunque no aparecen en el cuadro, en los casos de Argentina (1869 y 1895), Costa Rica (1904 y 1927) y Puerto Rico (1910 y 1920) se conservan censos más antiguos, cuya explotación permitiría el análisis de las transformaciones sociales, económicas y demográficas en el largo plazo (McCaa, Haines y Mulhare, 2000).

De los censos para los que se dispone de 100% de los casos se extraerán muestras de 10% de acuerdo con los procedimientos que se detallan más adelante. Asimismo, para los censos levantados entre 2000 y 2003 se extraerán muestras sistemáticas de diez por ciento.

En el caso de México, IPUMS-Internacional incluye muestras de los censos de 1960, 1970, 1990 y 2000. Sin embargo, para su incorporación en IPUMS-AL la muestra de 1% de 1990 se ampliará hasta 10% y para el año 2000 se extraerá una nueva muestra. Para este año IPUMS-Internacional cuenta con una muestra estratificada por conglomera-

CUADRO 2
Densidad y tamaño estimado de las muestras por país y ronda censal

	Densidad de las muestras de microdatos (porcentajes)					*Tamaño de las muestras de microdatos, registros individuales (miles)*				
	1960	*1970*	*1980*	*1990*	*2000*	*1960*	*1970*	*1980*	*1990*	*2000*
Argentina	3	2	2	100	100	500	469	559	3 262	3 700
Bolivia	.	100	.	100	100	.	461	.	642	830
Brasil	25	25	25	12	10	7 028	9 252	11 752	14 205	17 000
Chile	1	5	100	100	100	88	443	1 133	1 335	1 520
Colombia	2	100	100	100	100	350	1 989	2 643	3 275	4 000
Costa Rica	6	100	100	.	100	82	187	242	.	360
República Dominicana	7	7	8	100	100	203	272	476	761	840
Ecuador	3	17	100	100	100	136	924	835	965	1 260
El Salvador	1	5	.	100	100	26	176	.	512	630
Guatemala	5	5	5	100	100	210	290	302	833	1 270
Honduras	1	10	100	.	100	19	278	425	.	610
México	1.5	1	n.a.	100	100	503	483	.	8 028	10 100
Nicaragua	n.a.	10	.	100	.	.	189	.	436	.
Panamá	5	20	100	100	100	54	286	182	233	280
Paraguay	5	10	100	100	100	90	234	303	415	550
Perú	n.a.	n.a.	n.a.	100	100	.	.	.	2 205	2 710
Puerto Rico	10	3	7	6	6	235	81	224	211	234
Uruguay	5	100	100	100	.	128	279	296	316	.
Venezuela	2	100	100	30	100	132	1 060	1 452	1 802	2 420
Total						9 784	17 353	20 824	39 436	48 314

. = No hubo levantamiento censal en esta década.
n.a. = Microdatos incompletos o perdidos, pero los censos fueron levantados.

dos de 10% con base en las áreas de enumeración básicas (áreas geoestadísticas básicas o AGEB, según la terminología censal mexicana) y localidades. Este diseño muestral es útil para el análisis múltiple, aunque no ofrece la precisión que podríamos obtener con otro tipo de diseño. La nueva muestra se ajustará al diseño ya aplicado en otros censos y se conformará a partir de los datos derivados del cuestionario corto, disminuyendo así el número de variables respecto a la muestra existente. En el caso de 1970, recientemente ha aparecido una muestra olvidada de 3%. Si resulta posible rescatar y documentar esta muestra, también entrará al sistema de IPUMS-AL. Cada una de las muestras tendrá sus propias virtudes, de ahí que los usuarios podrán elegir la que más convenga a sus intereses de investigación, puesto que ambas estarán disponibles en IPUMS-AL.

IPUMS-AL tiene especial interés en difundir los datos más recientes. El financiamiento que este proyecto dedica a la obtención de las licencias de difusión de los microdatos espera auxiliar a los institutos de estadística en la tarea de asignar el personal necesario para extraer y procesar muestras de uso público para la ronda censal de 2000.

La parte derecha del cuadro 2 informa del tamaño de las muestras que se integrarán. El número total de casos disponibles entre todos los países parte de aproximadamente 10 millones en la ronda de 1960 y llega a casi 50 millones en la de 2000. Con todos los países y censos integrados, la base de datos completa incluirá aproximadamente 135 millones de casos.

Como hemos apuntado anteriormente, la importancia de los microdatos en América Latina no es sólo cuestión de tamaño, sino de contenido. Gracias a los mencionados esfuerzos de Cota y Celade, la mayoría de los países comparten un gran número de variables de carácter individual y por hogar.

Aspectos técnicos de la integración: desafíos y oportunidades

Diseño muestral

En muchos casos los datos proporcionados proceden de los archivos que en su día fueron utilizados en la preparación de los volúmenes del censo que se publicaron. Por lo tanto, se trata de archivos de datos de uso exclusivo de los institutos de estadística. De ellos se extraerán muestras autoponderadas de 10%. El diseño muestral que se apli-

cará busca el equilibrio entre la precisión de la muestra y el costo de oportunidad en su desarrollo.

La unidad muestral es el hogar, por lo que el número de observaciones independientes de cada archivo censal es el número de hogares y no el número de individuos. Esta estrategia tiene implicaciones en cuanto a la eficiencia final de la muestra. El error estándar en muestras por conglomerados de hogares depende del número de conglomerados muestreados y de la homogeneidad de las variables dentro de cada conglomerado (Hansen, Hurwitz y Madow, 1953). En el peor de los casos, con homogeneidad perfecta dentro de los conglomerados, el error estándar por variable es inversamente proporcional a la raíz cuadrada del número de conglomerados y no del número de individuos. Para las variables heterogéneas dentro de los conglomerados, tales como la edad y el sexo, establecer conglomerados tiene un efecto mínimo.

Para algunas muestras la pérdida de eficiencia asociada a un diseño por conglomerados se compensa con una estratificación proporcionalmente ponderada. En particular, desde 1960 la Oficina de Censos de Estados Unidos ha incrementado la utilización de diseños muestrales estratificados polietápicos. Dichos procedimientos pueden generar muestras autoponderadas con bajas proporciones de errores estándar, particularmente para los casos de etnicidad, tamaño del hogar, y condición de pertenencia a viviendas colectivas. Sin embargo IPUMS-AL no utiliza este procedimiento dadas las desventajas que presenta en relación con su complejidad y alto costo.

La forma en que se han organizado los datos en los censos de América Latina permitirá crear muestras de alta precisión a bajo costo. A diferencia de los censos recientes de Estados Unidos, donde los cuestionarios son enviados por correo, los de América Latina son levantados mediante la enumeración directa. En cada censo un entrevistador acudió de vivienda en vivienda para conversar personalmente con los residentes. De la utilización de este método resulta un producto adicional: los registros se ordenan de acuerdo con la secuencia de enumeración dentro de cada distrito o demarcación enumerativa. En la práctica esto significa que los datos se encuentran organizados geográficamente dentro de los distritos o demarcaciones correspondientes.

IPUMS-AL aprovechará esta condición organizativa de los datos para crear sus muestras sistemáticas de hogares. Dentro de cada distrito o área de enumeración se designa al azar un punto de inicio entre el

1 y el 10 y, a partir de éste se selecciona cada décimo hogar. Así, por ejemplo, si el punto inicial es 5, se incorporan a la muestra los hogares que aparecen en el 5º, 15º, 25º lugares hasta concluir con el distrito o demarcación correspondiente. Con esta estrategia se alcanza una estratificación geográfica muy fina, con ponderación proporcional. Asimismo, como las características económicas y sociales de los individuos están altamente correlacionadas en el espacio, la muestra resultante adquiere mayor precisión que una muestra aleatoria simple por hogares.

Igualmente IPUMS-AL generará muestras de individuos en unidades colectivas de manera independiente. El censo es prácticamente la única fuente que puede utilizarse para generar microdatos de unidades como prisiones, hospitales, asilos de ancianos, campamentos de viviendas móviles, y cuarteles militares. Debido a los efectos de la estratificación, el censo de los residentes en unidades colectivas se encuentra sujeto a errores estándar de gran magnitud si se les aplica la misma estrategia que a las personas que habitan en hogares particulares. La Oficina de Censos de Estados Unidos y otras agencias estadísticas afrontan este fenómeno mediante el muestreo de grandes unidades con carácter individual en lugar de hacerlo por hogar. Este procedimiento permite mantener la representatividad muestral y a la vez mejorar la eficiencia de la muestra al incrementar el número de observaciones independientes de las unidades colectivas.

La definición de unidad colectiva varía ampliamente entre los países. Siguiendo el ejemplo de IPUMS-USA, IPUMS-AL propone una definición homologada que pueda ser empleada de manera consistente en todos los censos. Se trata de una definición basada por completo en el tamaño de la unidad. Todas las unidades con más de 30 residentes serán clasificadas como colectivas o mayores.

Para elaborar las muestras entre las unidades mayores se designa aleatoriamente un punto de inicio entre 1 y 10 y, a partir de aquí, se selecciona cada décimo individuo. Este procedimiento se modifica cuando es posible identificar que un grupo de familiares vive dentro de una unidad mayor. Es interesante preservar las relaciones interfamiliares para poder estudiar aspectos como la fecundidad, el tipo de matrimonios y la composición familiar. Así pues, cuando se encuentra una unidad familiar dentro de una unidad mayor, la unidad familiar entera se considera como un punto muestral único. Con esta estrategia, los individuos sin relaciones familiares y los grupos familiares tendrán una probabilidad de 10% cada uno de ser inclui-

dos en la muestra. Tanto para los individuos como para las familias se construirán variables informando del tamaño y la composición de la unidad mayor a la que pertenecen.

Corrección de errores y reformateo de datos

Las tareas de corrección de errores y reformateo de los datos son realizadas sistemáticamente por un programa que explora la estructura de los registros, reformatea los datos, revisa la consistencia interna de la información, y corrige los errores.

La experiencia acumulada con el proyecto IPUMS-Internacional (Esteve y Sobek, 2003) nos hace esperar una gran variedad de irregularidades en los datos de América Latina. En los 17 censos internacionales que hemos procesado hasta la fecha, los problemas en el formato de los datos afectan sólo a una pequeña fracción de los casos; no obstante, todos los datos deben ser analizados sistemáticamente a fin de producir muestras limpias. Las tareas de limpieza demandan una inversión de tiempo superior a la que normalmente se prevé inicialmente. Las bases de datos más antiguas –aquellas que datan de las décadas de 1960 y 1970– generalmente plantean los mayores problemas.

Los archivos de datos originales se encuentran preservados en una amplia variedad de formatos: *i)* Los archivos *rectangulares* representan el formato más simple, con información geográfica, de vivienda, de hogares y de familias, repetida en cada registro individual. *ii)* En los archivos *jerárquicos* los microdatos tienen hasta cuatro tipos de registros entrelazados. En estos archivos cualquier irregularidad en la secuencia numérica de los tipos de registro afecta a la generalidad de los datos. *iii)* Los censos *vinculados* están organizados en múltiples tipos de registros almacenados en archivos separados, diseñados para vincularse entre sí por medio de números comunes de identificación (ID). Pequeñas imperfecciones en los números de identificación (ID) pueden causar problemas significativos. *iv)* Finalmente, en las muestras de *matriz invertida* se coloca cada variable en un archivo separado. Esta estructura de datos es optimizada para una rápida tabulación y depende de una secuencia de casos perfecta dentro de cada archivo. Por fortuna los archivos de matriz invertida se encuentran, aparentemente, en excelentes condiciones, y es poco probable que ocasionen serios problemas.

La estandarización de formatos implica la conversión de cada muestra en un formato simple de tipo jerárquico, compuesto por un registro de hogar seguido por los registros individuales de sus miembros. Con este sistema, toda la información geográfica y de la vivienda se repite en cada hogar.

Los institutos nacionales de estadística no siempre verifican que haya consistencia entre las distintas jerarquías en que se organizan los datos censales. Frecuentemente encontramos que las distribuciones marginales de las características individuales y de los hogares concuerdan con los resultados publicados; sin embargo cuando se analizan detalladamente estos archivos afloran inconsistencias entre los distintos tipos de registros que dificultan la construcción de las muestras de microdatos. Estas inconsistencias incluyen hogares con personas perdidas, personas sin información de hogares, y hogares mezclados. A pesar de que estas irregularidades nunca implican a muchos casos, deben ser resueltas. IPUMS-AL proporcionará toda la documentación generada durante el proceso de corrección de estas inconsistencias para informar al usuario final.

Las limitaciones de espacio nos impiden describir detalladamente la amplia variedad de problemas que se encuentran relacionados con el formato, y explicar las soluciones ideadas en cada caso pues todo censo es diferente de los demás. Generalmente para solucionar un problema se utiliza información contenida en el mismo censo, razón por la cual las soluciones varían en función del censo que se está trabajando.

Verificación de consistencia, edición de datos, y corrección de datos no especificados

Una vez solucionadas las cuestiones relacionadas con el formato, en la siguiente etapa se procede a verificar la consistencia interna de la base de datos, la imputación, y la corrección de datos no especificados. Para ello se aplican distintas pruebas con las que se verifica la consistencia interna de los datos y, por extensión, la calidad general de las muestras. Aunque los microdatos de América Latina cuentan con gran prestigio, muchas de las muestras nunca han sido verificadas ni "limpiadas". Entre las pruebas que se realizan ha de comprobarse que en todos los hogares haya una persona de referencia o cabeza del hogar, que no haya hogares con múltiples esposas o cónyuges

de la persona principal en países en donde la práctica de la poligamia no está legalmente reconocida, y que no haya registros duplicados. Igualmente se revisan las inconsistencias tanto entre los hogares como entre los individuos. Por ejemplo, se verifica que la condición laboral, el estado matrimonial, el nivel educativo, y la asistencia escolar guarden consistencia con la edad del individuo. Cuando los errores en los datos pueden ser identificados sin ninguna duda, se advierte mediante una nota que los datos son inconsistentes.

Una vez examinada la consistencia interna de los datos, los valores especificados o inconsistentes se imputan. En Estados Unidos los valores no especificados o inconsistentes son rutinariamente reemplazados mediante procedimientos de imputación probabilísticos o basados en la inferencia lógica. Por ejemplo, cuando el sexo no se especifica se puede inferir del sexo del cónyuge si la persona está casada o unida. Cualquier dato imputado es debidamente marcado para otorgar al usuario la libertad de utilizarlo o no.

Cuando los datos no especificados o inconsistentes no pueden ser reemplazados por medio de la edición lógica por computadora, se utilizan los procedimientos de asignación probabilística diseñados por la Oficina de Censos de Estados Unidos. Para cada variable se cuenta con una serie de criterios para imputar la información. Estos criterios se establecen mediante el análisis de los mejores pronosticadores de cada variable, y pueden ser diferentes de censo a censo. Por ejemplo, si la información sobre asistencia escolar no se especifica, es posible imputar este dato tomando como base la del individuo más cercano en el archivo que comparte la misma edad, sexo, grupo étnico y estatus socioeconómico de los padres. Cuando no se puede encontrar un "donante" del todo compatible, se utiliza el registro que cumple con la mayor cantidad de criterios. El valor "donante" está sujeto a verificaciones de consistencia y es rechazado si se considera inapropiado. Una señal de calidad de datos identifica los datos reemplazados.

Si bien el reemplazo de los datos no especificados o inconsistentes mejora significativamente la confiabilidad de la estimación muestral y simplifica el uso de las muestras, no es habitual aplicar esta técnica fuera de Estados Unidos. Dada la experiencia adquirida en proyectos anteriores en la aplicación de estos métodos, creemos oportuno aplicarlos también, cuando sea necesario, a los datos de América Latina. Todos los cambios realizados estarán completamente documentados y el usuario, si así lo requiere, podrá prescindir de los datos que han sido modificados.

Armonización

La armonización de los datos de América Latina se hará con base en el trabajo de armonización desarrollado en IPUMS-Internacional. Las muestras censales internacionales emplean diferentes sistemas de clasificación numéricos, cuya conciliación es un aspecto de suma importancia para este proyecto. El diseño de las variables influye frecuentemente en las estrategias analíticas que adoptan los investigadores.

La Organización de las Naciones Unidas cuenta con dos proyectos a gran escala de armonización regional de microdatos censales. El primero de ellos fue el proyecto Omuece, descrito anteriormente. Dentro de este proyecto Celade creó versiones estandarizadas para 29 censos latinoamericanos que fueron levantados entre 1960 y 1976 (McCaa y Jaspers, 2000). El segundo proyecto lo desarrolló el Population Activities Unit de las Naciones Unidas (PAU) en Génova (Botev, 2000); actualmente está en curso y persigue la estandarización de muestras de microdatos de las rondas censales de 1990 y 2000 de 16 países de Europa y Norteamérica. Estas dos iniciativas han proporcionado a IPUMS-Internacional valiosa información sobre cómo afrontar el reto de la integración.

Los dos proyectos de la ONU se apoyan en distintas filosofías en su diseño. Omuece incluyó sólo las variables que estaban presentes en todos los censos y lo hizo con base en su mínimo común denominador; la mitad de las variables quedaron excluidas y se perdió el detalle de la codificación original de las que fueron integradas. La pérdida del detalle afectó tan severamente las bases de datos que la mayoría de los usuarios optó por trabajar con las muestras originales, aun siendo incompatibles. El proyecto PAU representa el extremo opuesto. En este caso no existe ningún interés por estandarizar los esquemas de códigos para variables categóricas complejas tales como religión, relaciones familiares, ocupación, grupo étnico, o lengua. Sólo las variables más simples, tales como edad, sexo, estado matrimonial, y relación con la actividad, son recodificadas dentro de un esquema común. Las transformaciones de datos dentro del proyecto PAU logran hacer más sencillas las comparaciones internacionales, pero todavía no están concluidas.

La estrategia diseñada por IPUMS-Internacional consiguió superar con éxito los problemas asociados a las dos alternativas anteriores. A diferencia de Omuece, IPUMS-Internacional mantiene todos los detalles provistos en las muestras originales. A diferencia de PAU, IPUMS-

Internacional ofrece datos completamente integrados. Para lograr estos objetivos se emplean distintas estrategias: en algunos casos las variables originales son compatibles y recodificarlas dentro de una clasificación común es algo sencillo, sin embargo la mayoría de las variables no permiten una clasificación uniforme simple sin que se pierda información. Para un mismo concepto algunos censos proveen más información que otros, por lo que la aplicación del mínimo común denominador acarrearía la pérdida de detalles; en estos casos se construyen esquemas de codificación múltiples, compuestos de varios dígitos que informan de los distintos rangos de la variable. El primero o el segundo dígito de cada código ofrece información disponible en todas las muestras. El tercero o cuarto dígito añade información adicional que suele estar presente en la mayoría de censos. Finalmente, los últimos dígitos informan de detalles disponibles en un número reducido de muestras.

Más allá de la adecuación de los datos de América Latina al estándar de IPUMS-Internacional, IPUMS-AL desarrollará sus propios estándares, mejor adaptados a las necesidades y características de las variables en esta región. El usuario podrá elegir el tipo de clasificación que desee. El esquema de clasificación de la variable estado matrimonial sirve para ilustrar este punto (véase el cuadro 3). Conforme al diseño de IPUMS-Internacional, el primer dígito de estado matrimonial tiene cuatro categorías comparables en todos los censos: soltero, casado/unido, separado/divorciado/esposo(a) ausente, y viudo(a). Como la distinción entre divorciado(a) y separado(a) no se mantiene en todas las muestras, no es posible introducir esta diferenciación en el primer dígito de la variable. Así pues, el segundo dígito distingue a los divorciados de los separados y a los casados formalmente de los unidos consensualmente. El tercero y último dígito establece la diferencia entre tipos de matrimonios (civil, religioso, poligámico), información disponible sólo en pocos países.

Todas las muestras en América Latina distinguen claramente las uniones libres de los matrimonios legales o civiles, razón por la cual la versión de la variable estado matrimonial para América Latina incluirá en su primer dígito un código para las uniones libres. El sistema de acceso a los datos recomendará, por defecto, esta versión de la variable, a no ser que el usuario haya especificado previamente su interés por comparar datos entre regiones.

Las variables geográficas plantean los mayores retos. IPUMS-AL no pretende lograr la armonización completa en la información geográfi-

CUADRO 3

Clasificados de la variable estado matrimonial y disponibilidad de categorías

		Colombia				*Francia*					*Kenia*		*México*				*Estados Unidos*				*Vietnam*	
Código	*Etiqueta*	*64*	*73*	*85*	*93*	*62*	*68*	*75*	*82*	*90*	*89*	*99*	*60*	*70*	*90*	*00*	*60*	*70*	*80*	*90*	*89*	*99*
100	Solteros/nunca unidos	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	Casados/en unión																					
210	Matrimonio																					
	(sin especificar)	X	X	X	X	X	X	X	X	X	X	X	.	.	.	.	X	X	X	X	X	X
211	Matrimonio civil	.	.	.	.	.	.	.	.	.	.	.	X	X	X	X	.	.	.	.	.	.
212	Matrimonio religioso	.	.	.	.	.	.	.	.	.	.	.	X	X	X	X	.	.	.	.	.	.
213	Matrimonio civil																					
	y religioso	.	.	.	.	.	.	.	.	.	.	.	X	X	X	X	.	.	.	.	.	.
214	Matrimonio poligámico	.	.	.	.	.	.	.	.	.	X	X	.	.	.	.	.	.	.	.	.	.
220	Unión libre	X	X	X	X	.	.	.	.	.	.	.	X	X	X	X	.	.	.	.	.	.
	Separados/divorciados/ cónyuge ausente																					
310	Separados o divorciados	.	X	X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
320	Separados	X	.	.	.	.	.	.	.	.	X	X	.	X	X	X	X	X	X	X	X	X
330	Divorciados	.	.	.	.	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
340	Casados, cónyuge ausente																					
	(no especificado)	X	X	X	X	X	X	X	X	X	X	X	.	.	.	.	X	X	X	X	X	X
341	Matrimonio civil	.	.	.	.	.	.	.	.	.	.	.	X	X	X	X	.	.	.	.	.	.
342	Matrimonio religioso	.	.	.	.	.	.	.	.	.	.	.	X	X	X	X	.	.	.	.	.	.
343	Matrimonio civil																					
	y religioso	.	.	.	.	.	.	.	.	.	.	.	X	X	X	X	.	.	.	.	.	.
344	Matrimonio poligámico	.	.	.	.	.	.	.	.	.	X	X	.	.	.	.	.	.	.	.	.	.
350	Unión libre,																					
	cónyuge ausente	X	X	X	X	.	.	.	.	.	.	.	X	X	X	X	.	.	.	.	.	.
400	Viudos	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

X = Categoría disponible en la muestra.

Nota: Las muestras están identificadas con los dos últimos dígitos del año censal que representan.

ca más específica, pero intentará crear una definición integrada de áreas metropolitanas. Siempre que sea posible, IPUMS-AL proporcionará las bases cartográficas para todas las escalas disponibles en los datos.

Se estima que el proceso de armonización requerirá en total aproximadamente 850 000 transformaciones de datos. Cada transformación debe ser planeada, ejecutada, verificada, vuelta a verificar, y documentada. Este trabajo representa casi un tercio del esfuerzo requerido para el proyecto.

Variables construidas

IPUMS-AL construirá nuevas variables para mejorar la utilidad de los datos. Algunas de ellas son muy simples, tales como el número de serie, año censal, código de país, tamaño de la unidad, y factor de ponderación. Otras son más complicadas.

Las autoridades censales de América Latina colectan datos sobre los hogares y las relaciones que se establecen entre los individuos de un mismo hogar. A partir de esta información se crean variables de carácter individual que dan la posición dentro del hogar de la madre, el padre y el cónyuge (o pareja) del individuo; tales indicadores figuran entre las mayores contribuciones que pueden hacerse a las bases de datos. Este tipo de variables permite, por ejemplo, contrastar fácilmente las características de dos personas unidas en matrimonio o en unión libre.

También se crean variables integradas que informan de las características del hogar y la familia en forma individual. Algunos de estos indicadores –tales como pertenencia a un grupo familiar, tamaño de la familia, número de hijos propios, número de hijos propios menores de cinco años de edad, y edad de los hijos propios mayor y menor– se encuentran ya incorporadas en IPUMS-Internacional.

Finalmente IPUMS-AL abordará la construcción de variables que describan el estatus socioeconómico. Relativamente pocos censos de América Latina dan información directa sobre el ingreso económico, por lo cual la ocupación y las características del hogar son probablemente los mejores indicadores para estimar el estatus socioeconómico. En el caso de IPUMS-USA se diseñaron dos medidas basadas en la ocupación para reflejar el estatus socioeconómico: índice socioeconómico de Duncan y el nivel de ingreso económico. Los investigadores han usado ambas medidas de forma extensiva (Sobek, 1995, 1996, 1997;

Treiman, 1977; Nakao y Treas, 1992; Ganzeboom y Treiman, 1996; Ganzeboom, De Graaf y Treiman, 1992). En América Latina se está trabajando con nuevos indicadores socioeconómicos basados en la información sobre ocupación y hogar.

Documentación

La creación de una documentación integrada y comprensible es un componente central del proyecto, pero también uno de sus más grandes retos. Afortunadamente IPUMS-AL cuenta con una colección significativa de material censal original. Con el soporte económico de la subvención otorgada a IPUMS-Internacional, Celade ha inventariado, catalogado y escaneado un amplio rango de documentos de los censos de América Latina. Además, el Minnesota Population Center es el depositario del archivo histórico de documentos censales de la División de Estadística de las Naciones Unidas, gracias a la donación concedida por esta institución. Finalmente, la tercera fuente de documentación y soporte técnico proviene directamente de los institutos de estadística de cada país y de los expertos nacionales contratados para asesorar las tareas de la integración.

La documentación integrada cubre una gran cantidad de aspectos: procedimientos e instrucciones de enumeración, corrección de errores y otros procesamientos postenumerativos, diseños muestrales, cuestionarios, y análisis de calidad de los datos. Celade facilitará las traducciones de los documentos integrados que sean más relevantes para el proyecto.

Para cada variable se proporciona una descripción detallada que incluye definiciones del universo, distribución de frecuencias y código de variables. La descripción de las variables más importantes se complementa con una serie de acotaciones sobre su comparabilidad, que alertan sobre las discrepancias que puedan existir entre una determinada muestra y el esquema general de clasificación.

Asimismo, en la documentación se describen todas aquellas transformaciones que se efectuaron en los datos originales a fin de generar la base integrada. Como no se pierde ningún detalle de la información original, el usuario puede deshacer todas esas transformaciones si desea disponer de la información original.

Se estima que las series de datos requerirán aproximadamente un millar de páginas de documentación. Para manejar tal cantidad de información, el sistema de acceso a los metadatos, con base en

Internet, mostrará sólo la documentación que concierne a los criterios que el usuario irá especificando durante el proceso de selección. Por ejemplo, si un usuario selecciona sólo los censos de Venezuela, recibirá exclusivamente la información relativa a las muestras de ese país. Cuando todas las muestras estén disponibles en Internet, la habilidad del sistema para filtrar sólo la información relacionada con cada solicitud será un elemento clave para la ágil navegación en el sistema de extracción.

Difusión

El acceso a los datos es un aspecto esencial del proyecto. La difusión de los datos debe ser altamente efectiva para optimizar su utilización. La complejidad de la nueva base de datos será más grande que la de cualquiera que se haya desarrollado previamente en el marco de IPUMS, pero IPUMS-AL prevé simplificar aún más el acceso a los microdatos y metadatos.

IPUMS viene trabajando en métodos de diseminación electrónica para datos y documentación en ciencias sociales desde hace 10 años; ha desarrollado el más poderoso sistema de extracción de datos por Internet. El proyecto IPUMS-USA fue pionero en la divulgación y distribución de datos a gran escala y ha inspirado otras iniciativas en el ámbito de las ciencias sociales. IPUMS-Internacional está desarrollando en la actualidad una segunda generación de *software* de diseminación de datos. El nuevo sistema de acceso a datos proporcionará herramientas avanzadas para la consulta de documentación, cuerpos de datos definidos, construcción de variables específicas, y adición de información de contexto.

Dado que las series latinoamericanas de datos incorporarán más de cien millones de observaciones y cientos de variables provenientes de docenas de censos, la habilidad para unir y crear divisiones de datos resulta crítica. En este sentido IPUMS-AL se beneficiará de todas las innovaciones que se produzcan en el contexto de IPUMS-Internacional.

La maquinaria de extracción de datos está diseñada para tomar entera ventaja de la estructura jerárquica de los datos censales. Los investigadores tienen la opción de obtener los datos en formato rectangular o jerárquico y la posibilidad de solicitar hogares completos con base en las características de uno de sus integrantes. Por ejemplo, podrán requerir aquellos hogares en los que residan personas mayo-

res de 90 años. Las versiones futuras del sistema de acceso de datos de IPUMS-Internacional prevén agregar dos características adicionales para simplificar la explotación de la estructura jerárquica de los datos.

1) Un procedimiento para anexar las características de las cabezas de hogar, cabezas de familia, cónyuges, madres y padres a cada registro individual. Por ejemplo, el sistema permitirá a los estudiosos del matrimonio crear nuevas variables que describan la edad del cónyuge o su lugar de nacimiento.
2) Un procedimiento para contar el número de personas dentro de cada hogar, familia, o hijos propios para cada padre que tenga una combinación de hasta cuatro características. Por ejemplo, el sistema de acceso de datos será capaz de contar el número de hijas adolescentes en el mercado laboral para cada madre con hijos que viven en el mismo domicilio. El sistema también adicionará elementos numéricos (por ejemplo, ingresos económicos) propios de los hogares, las familias o los hijos propios.

Finalmente, IPUMS-AL ofrecerá a los usuarios la posibilidad de replicar los extractos de datos que se han usado en algunos estudios publicados. La habilidad para replicar estudios existentes es esencial para el desarrollo científico. El nuevo sistema que se está desarrollando para IPUMS-Internacional prevé la entrega con cada extracto de un identificador. Una vez publicados los resultados, el sistema solicitará a los usuarios que introduzcan el número de identificador del extracto utilizado en su investigación. De esta manera, si un usuario quiere disponer de los mismos datos utilizará el número de identificación correspondiente. Junto con los datos recibirá un documento con las citas de todas las publicaciones que utilizaron esa misma información.

Aplicaciones de investigación: algunos ejemplos

Todos los esfuerzos realizados en la integración y difusión de las muestras de microdatos censales de América Latina están dirigidos principalmente a maximizar la utilización de los datos, pues estamos convencidos de su potencial. IPUMS-AL espera influir fuertemente en las ciencias sociales, pues abre un océano de oportunidades para los investigadores. A continuación se exponen algunas de las potenciales aplicaciones de estos datos.

Envejecimiento

Las muestras de microdatos censales de América Latina constituirán un recurso de gran importancia para el estudio de la población en edad avanzada. Gracias a la cobertura histórica de los datos será posible realizar un análisis por cohortes (Palloni, 2002; Chackiel, 2001; Viveros Madariaga, 2001). Además, para el desarrollo de los nuevos métodos de proyección de la población anciana se requieren múltiples parámetros que pueden ser obtenidos con mayor facilidad a partir de grandes muestras de microdatos (véase por ejemplo, Vaupel, Yi y Zhenglian, 1997). Sin duda alguna es importante que IPUMS-AL logre brindar nuevas oportunidades para realizar investigaciones comparativas entre las naciones sobre el envejecimiento. Este tipo de estudios comparativos son valiosos elementos que en otras regiones del mundo sirven para tomar decisiones políticas (Gruber y Wise, 1998, 1999; Johnson, 1999; Hermalin y Chan, 2000).

Migración

En las décadas recientes América Latina se ha convertido en una región de emigración neta y Estados Unidos en su principal lugar de destino (De Launey y Tapinos, 2001; Canales Cerón, 2001). A partir del decenio de 1980 muchos censos latinoamericanos empezaron a captar información sobre el número de hijos de cada hogar que residen fuera del país de origen. Las preguntas retrospectivas sobre migración derivan de un fuerte interés por los movimientos hacia y desde Estados Unidos. La utilización de los datos de IPUMS-AL junto con los de IPUMS-USA permitirá contrastar las características de los individuos que residen en un país latinoamericano respecto a las de quienes se encuentran en Estados Unidos y provienen de ese mismo país. La estructura jerárquica de los datos facilita el estudio de los individuos en sus contextos familiares y el conocimiento de su hogar, y hace posible, por ejemplo, investigar las características de los familiares de quienes son padres o madres solteros latinoamericanos, y residen en Estados Unidos o en América Latina.

Fecundidad

De 1960 a 2001 la tasa global de fecundidad para América Latina descendió de un promedio superior a 6 niños por mujer a 2.8. Esta acelerada transición se ha convertido en un prometedor y fructífero tema de interés para los académicos (Guzmán *et al.*, 1996). IPUMS-AL facilitará el estudio de los patrones diferenciales de fecundidad por grupos ocupacionales, región, educación, tamaño de localidad, y una multitud de variables adicionales de carácter individual, familiar o comunitario. La riqueza de estos datos mejorará sustancialmente las posibilidades de análisis de los determinantes del descenso de la fecundidad en los países en desarrollo. Desde la década de 1970, los censos latinoamericanos han consignado regularmente el número de hijos nacidos vivos y de hijos sobrevivientes, la fecha de nacimiento del último hijo nacido y la condición de supervivencia para las mujeres en edad fértil. Adicionalmente, las series de microdatos incorporarán un conjunto de vínculos entre madres e hijos y facilitarán el análisis de la fecundidad por el método de los *hijos propios*.

Salud pública

Los censos latinoamericanos han captado históricamente información relacionada con la salud pública, como la disponibilidad de servicios sanitarios, la fuente de suministro de agua, el tipo de combustible empleado para cocinar, y los materiales de construcción de las viviendas (De Vos y Arias, 1996). Al complementarlos con variables relativas a la supervivencia infantil y la mortalidad, estos datos ofrecerán oportunidades excepcionales para estimar las condiciones de salud pública locales, regionales y nacionales.

Análisis comparativo de políticas públicas

La disponibilidad de microdatos altamente comparables entre países con distintas políticas públicas brinda un excelente banco de pruebas para medir su eficacia. En Estados Unidos esta estrategia ha sido una herramienta muy útil para estimar los efectos de las variaciones interestatales en programas de asistencia social, acceso a los servicios de salud, y políticas fiscales (véase por ejemplo Duncan y Hoffman,

1992; Lundbert y Plotnik, 1995; Moffitt, 1992; Ruggles, 1997; Whittington, 1993). Los mismos modelos pueden ser aplicados a los países de América Latina también para estimar el efecto de las políticas públicas en el desarrollo económico, la desigualdad social, la urbanización, y el cambio demográfico.

Los ejemplos citados aquí son sólo una muestra de las aplicaciones de la nueva base de datos. Evidentemente existen muchas más posibilidades de uso en campos como la demografía de la violencia, las consecuencias sociales de las discapacidades físicas, los cambios en la estructura familiar, las transformaciones en la estructura ocupacional, la urbanización, la migración interna, el trabajo infantil, la nupcialidad, la educación, la universalización de la enseñanza pública, la participación femenina en la actividad económica (McCaa *et al.*, 2000, 2003) y un largo etcétera.

Con el esfuerzo de todos, los tesoros estadísticos de América Latina cuidadosamente preservados por Celade, estarán a disposición de la comunidad científica internacional en un plazo de cinco años. Sin el pertinaz empeño de varias instituciones y de demógrafos como el doctor Gustavo Cabrera, hoy el acceso a estos datos sería imposible. El proyecto IPUMS para América Latina asume con responsabilidad la tarea de construir la más importante fuente de información para el estudio de las sociedades latinoamericanas. Para ello cuenta con la inestimable cooperación de instituciones como Celade, los institutos nacionales de estadística, y los más destacados expertos nacionales. Aunque el éxito final del proyecto está en manos de los investigadores y depende de la capacidad de la base de datos para cumplir con las expectativas que se han generado.

A juzgar por la experiencia de México, los resultados que obtendremos para el conjunto de los países de América Latina son muy esperanzadores. La cooperación entre las distintas partes que colaboran en la integración de los microdatos mexicanos –el INEGI, los expertos, y el Minnesota Population Center– ha dado sus frutos. La estrategia de armonización de IPUMS-Internacional ha permitido acomodar los censos de México al diseño global de integración sin perder detalle alguno de su idiosincrasia y riqueza conceptual. Desde su puesta en marcha en mayo de 2002, IPUMS-Internacional ha recibido un gran número de solicitudes para utilizar los datos mexicanos. Estos datos están siendo aplicados en una amplia gama de temáticas, como la migración de retorno, las pautas de nupcialidad, la participación femenina en el mercado de trabajo, los hogares con personas ancianas, la

escolarización y el trabajo infantil, el desarrollo económico, la pobreza y el descenso de la fecundidad, para citar sólo algunos ejemplos. Asimismo los datos de México se utilizarán en estudios comparativos entre países. La petición simultánea de datos de Estados Unidos y México con el objetivo de comparar la situación de los mexicanos en ambos países es recurrente en muchas de las solicitudes que implican datos mexicanos. Hasta la fecha la respuesta de los usuarios ha sido positiva. Apenas hemos recibido reparos a los esquemas propuestos para clasificar las variables. Los usuarios, con sus estudios, están día a día legitimando la base de datos, haciendo de ella un catalizador de investigaciones cada vez más ambiciosas en sus objetivos.

Bibliografía

Botev, Nikolai (2000), "PAU Census Microdata Samples Project", en Patricia Kelly Hall, Robert McCaa y Gunnar Thorvaldsen (coords.), *Handbook of International Historical Microdata for Population Research*, Minneapolis, Minnesota Population Center, pp. 303-317.

Canales Cerón, Alejandro I. (2001), "Factores demográficos del asentamiento y la circularidad en la migración México-Estados Unidos", *Notas de Población,* núm. 28, pp. 123-158.

Celade (1989), "El contenido demográfico de la boleta de los censos de población de la década del 90", en *Censos de Población de 1990: selección de documentos del Celade,* Santiago (Serie A-Celade, 193).

Chackiel, Juan (2001), "El envejecimiento de la población latinoamericana", en Rolando Franco (coord.), *Sociología del desarrollo, políticas sociales y democracia: estudios en homenaje a Aldo E. Solari*, México, Siglo XXI/CEPAL, pp. 166-185.

Delaunay, Daniel y George Tapinos (2001), "¿Se puede hablar realmente de la globalización de los flujos migratorios?", *Notas de Población,* núm. 73, pp. 15-49.

De Vos, Susan y Elizabeth Arias (1996), "Using Housing Items to Indicate Socioeconomic Status: Latin America", *Social Indicators Research,* núm. 38, pp. 53-80.

Dirección General de Estadística (1972), *IX Censo general de población, 1970. Resumen General*, México, Dirección General de Estadística.

——— (1962), *VIII Censo general de población, 1960. Resumen general*, México, Dirección General de Estadística.

Duncan, Greg J. y Saul D. Hoffman (1992), "Welfare Benefits, Economic Opportunities, and Out-of-Wedlock Births among Black Teenage Girls", *Demography,* núm. 27, pp. 519-535.

Esteve, A. y M. Sobek (2003), "Challenges and Methods of International Census Harmonization", *Historical Methods,* núm. 36, pp. 66-79.

Eternod, Marcela y Juan María Trejo (2001), "Homologación de las características económicas de la población en los censos mexicanos", documento preliminar, IPUMS-Internacional.

Ganzeboom, Harry y Donald Treiman (1996), "Internationally Comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations", *Social Science Research,* núm. 25, pp. 201-239.

———, P. De Graaf y Donald Treiman (1992), "A Standard International Socio-Economic Index of Occupational Status", *Social Science Research,* núm. 21, pp. 1-56.

Gruber, Jonathan y David A. Wise (1999), *Social Security and Retirement Around the World*, Chicago, University of Chicago Press.

——— y David A. Wise (1998), "Social Security and Retirement: An International Comparison", *American Economic Review Papers and Proceedings,* núm. 88, pp. 158-163.

Guzmán, José Miguel, Susheela Singh, Germán Rodríguez y Edith A. Pantelides (1996), *The Fertility Transition in Latin America,* Oxford, Clarendon Press.

Hall, Patricia Kelly, Robert McCaa y Gunnar Thorvaldsen (2000), *Handbook of International Historical Microdata for Population Research,* Minneapolis, Minnesota Population Center.

Hansen, Morris, William Hurwitz y William Madow (1953), *Sample Survey Methods and Theory,* Nueva York, Wiley.

Hermalin, Albert I. y A. Chan (2000), "Work and Retirement among the Older Population in Four Asian Countries: A Comparative Analysis", *CAS Research Paper Series*, núm. 22, Singapur, Center for Advanced Studies, National University of Singapore.

INEGI (2000), *XII Censo general de población y vivienda, 2000,* México, Instituto Nacional de Estadística, Geografía e Informática, en www.inegi.gob.mx.

——— (1993), *Resumen general. Resultados definitivos. Tabulados básicos. XI Censo general de población y vivienda, 1990*, México, Instituto Nacional de Estadísticas, Geografía e Informática.

IPUMS-Internacional (2000), *Muestras censales integradas, 1960, 1970, 1990 y 2000*, Minnesota, Minnesota Population Center (Datos y Documentos, CD).

Johnson, Paul (1999), *Pension Provision and Pensioners' Incomes in Ten OECD Countries,* Londres, Institute for Fiscal Studies.

Lundberg, Shelley y Robert A. Plotnik (1995), "Adolescent Premarital Childbearing: Do Economic Incentives Matter?", *Journal of Labor Economics,* núm. 13, pp. 177-200.

McCaa, Robert, Albert Esteve, Rodolfo Gutiérrez y Gabriela Vásquez (2003), "Women in the Workforce: Calibrating Census Microdata Against Gold Standards Mexico, 1990-2000", *Population Association of America Annual Meeting*, Minneapolis.

———, Rodolfo Gutierréz y Gabriela Vásquez (2000), "La mujer mexicana económicamente activa: ¿son confiables los microdatos censales? Una prueba a través de censos y encuestas. México y los Estados Unidos, 1970-1990", *Papeles de Población,* vol. 6, núm. 25, pp. 151-178.

———, Michael R. Haines y Eileen M. Mulhare (2000), "Argentina: First with Public Historical Census Microdata", en Patricia Kelly Hall, Robert McCaa y Gunnar Thorvaldsen (coords.), *Handbook of International Historical Microdata for Population Research*, Minneapolis, Minnesota Population Center, pp. 13-22.

——— y Dirk J. Jaspers-Faijer (2000), "The Standardized Census Sample Operation (Omuece) of Latin America, 1959-1982 [1995]: a Project of the Latin American Demographic Center (Celade)", en Patricia Kelly Hall, Robert McCaa y Gunnar Thorvaldsen (coords.), *Handbook of International Historical Microdata for Population Research*, Minneapolis, Minnesota Population Center, pp. 287-302.

——— y Steven Ruggles (2002), "The Census in Global Perspective and the Coming Microdata Revolution", en J. Carling (coord.), *Nordic Demography: Trends and Differentials, Scandinavian Population Studies*, vol. 13, Oslo, Unipub/Nordic Demographic Society, pp. 7-30.

Moffitt, Robert (1992), "Incentive Effects of the U.S. Welfare System: A Review", *Journal of Economic Literature,* núm. 30, pp. 1-61.

Nakao, Keiko y Judith Treas (1992), "The 1989 Socioeconomic Index of Occupations: Construction from the 1989 Occupational Prestige Scores", *GSS Methodological Report*, núm. 74, Chicago, National Opinion Research Center.

Palloni, Alberto (2002), "Demographic and Health Conditions of Aging in Latin America and the Caribbean", *International Journal of Epidemiology,* núm. 31, pp. 762-771.

Rabell, Cecilia (2000), "Mexico-Census Microdata: 1960, 1970, 1990, 1995", en Patricia Kelly Hall, Robert McCaa y Gunnar Thorvalden (coords.), *Handbook of International Historical Microdata for Population Research*, Minneapolis, Minnesota Population Center.

Ruggles, Steven (1997), "The Effects of AFDC on American Family Structure, 1940-1990", *Journal of Family History,* núm. 22, pp. 307-325.

——— y Matthew Sobek *et al.* (1997), *Integrated Public Use Microdata Series: Version 2.0*, Minneapolis, Historical Census Projects, University of Minnesota.

Sobek, Matthew (1997), *A Century of Work: Gender, Labor Force Participation, and Occupational Attainment in the United States, 1880-1990*, tesis de doctorado, University of Minnesota.

——— (1996), "Work, Status and Income: Men in the American Occupational Structure Since the Nineteenth Century", *Social Science History,* núm. 20, pp. 169-207.

——— (1995), "The Comparability of Occupations and the Generation of Income Scores", *Historical Methods,* núm. 28, pp. 47-51.

Treiman, Donald (1977), *Occupational Prestige in Comparative Perspective,* Nueva York, Academic Press.

Vaupel, James, Zeng Yi y Wang Zhenglian (1997), "A Multi-Dimensional Model for Projecting Family Households with an Illustrative Numerical Application", *Mathematical Population Studies,* núm. 6, pp. 187-216.

Viveros Madariaga, Alberto (2001), *Envejecimiento y vejez en América Latina y el Caribe: políticas públicas y las acciones de la sociedad*, Santiago, CEPAL (Población y Desarrollo, 22).

Whittington, Leslie A. (1993), "State Income Tax Policy and Family Size: Fertility and the Dependency Exemption", *Public Finance Quarterly*, núm. 21, pp. 378-398.