Retos técnicos

y oportunidades del análisis secundario de datos

□ Guillermo Zúñiga

El análisis secundario de datos consiste en la utilización de fuentes de datos que originalmente se capturaron con un propósito diferente al del usuario actual. Un ejemplo clásico es el de los datos del Instituto Nacional de Estadística Geografía e Informática (INEGI), o los datos sobre criminalidad nacional y de otros países. Los datos generados por los investigadores, que han recibido financiamiento por parte de organismos nacionales o internacionales, suelen dar como resultado bases de datos con diversas posibilidades de análisis, los cuales van más allá de las preguntas de investigación e hipótesis de los autores originales.

Otra importante ventaja de los datos secundarios es que si se utilizan con otro propósito, se elimina la influencia de las predisposiciones escolásticas, teóricas y propias de la disciplina de investigación original, ya que éstas, en todo caso, ejercieron su influencia de predisposición en el cuestionario original. Asimismo, se elimina la influencia de las respuestas complacientes prototípicas del campo de inves-

tigación, al utilizarse con otro propósito analítico, haciendo posible la respuesta a interrogantes que de otro modo no se podrían contestar.

Por ejemplo, el autor de estas líneas no encontró la riqueza de datos sobre delito en una encuesta diseñada sobre criminalidad, en comparación con las posibilidades analíticas sobre delito de la Encuesta Nacional de la Juventud 2005, que indagó la situación de los jóvenes en el país. Sucede que los investigadores o institutos que elaboran instrumentos de recopilación de datos reaccionan a diversas fuentes objetivas (marcos teóricos) y subjetivas (compromisos personales), que finalmente establecen lo que se va a explorar en el instrumento y la manera en que se hará.

A pesar de estas ventajas, es muy improbable que las bases de datos originales tengan todas las variables que el investigador requiere para dar respuesta a su modelo teórico o a las hipótesis base. Este tipo de carencia puede solventarse al agregar las variables requeridas a partir de otras fuentes. Esto se logra, ya sea escribiendo los valores para cada caso, cuya tarea es casi imposible si se trata de los datos de muchas personas, o utilizando las posibilidades de programación del propio programa estadístico (SPSS o SAS). También se emplea programación externa de paquetes que manejan datos, como el Visual Basic o el Fox Pro.

De hecho, trabajar con bases de datos secundarios no debe limitarse al uso de las variables tal cual se encuentran en la base de datos original, puesto que el nivel para lograr un grado de maestría o de doctorado exige un trabajo intelectual original. Aun así, siempre será necesario efectuar adiciones o modificaciones a la base de datos original, para llegar a las respuestas necesarias a nuevos problemas de investigación. Tampoco se hace justicia a la enorme cantidad de información científica, actualmente disponible, en términos del dinero que se invirtió, ni del esfuerzo material e intelectual. Cada investigador, en especial el que se inicia, quisiera hacer aportaciones originales con sus propios datos, pero a partir de datos secundarios un estudiante puede realizar incluso una investigación más original y abordar objetos de estudio innovadores. Es evidente el ahorro monetario en el mismo proceso de aplicación, si se trabaja con fuentes secundarias de datos. Sin embargo, el tiempo y esfuerzo invertidos en la investigación total son los mismos si se trabaja con datos secundarios que con datos propios.

Presentaremos enseguida un ejemplo de investigación de análisis secundario de datos en la que, con programación, se encontró una catalogación diferente de los tipos de familia en México.

Los tipos de familia

Una de las múltiples áreas de oportunidad analítica adicional de una base de datos obtenida son los tipos de estructura familiar en México. La comunidad científica y las polémicas y debates sobre los tipos de familia han concluido que la estructura familiar es demasiado diversa y compleja. La clasificación de los tipos de familia utilizada en la Encuesta Nacional de Ingresos y Gastos en los Hogares (ENIGH) sigue la clasificación estándar del INEGI: hogares nucleares, ampliados, compuestos, unipersonales y de corresidentes. Sin embargo, es posible efectuar otro tipo de análisis de los grupos familiares en la ENIGH.

El problema técnico

Más a detalle, en cuanto a la relación familiar sostenida en los grupos familiares, se puede efectuar la identificación del jefe o jefa de la familia con base en la metodología de aplicación empleada en la ENIGH. En la metodología de aplicación se establece: "El informante adecuado de la encuesta puede ser el jefe (a) del hogar, esposa o compañera, o un integrante del hogar de 15 años o más. La primera pregunta al informante adecuado al iniciar la aplicación fue:"¿Cuál es el nombre de los integrantes de este hogar empezando por el jefe o jefa?", se efectúa así la identificación del jefe o jefa de la familia. Esto es, el informante adecuado de la familia, quien quiera que haya sido, efectuó la decisión de la jefatura de la familia (padre, madre o algún otro pariente) al contestar la pregunta. Lo importante es que la primera

persona de cada grupo familiar siempre será el jefe o jefa de la familia, y la segunda persona, si existe, siempre será el compañero o compañera del jefe o jefa de familia.

La elección del jefe o jefa de familia por parte del informante adecuado se efectuó bajo condiciones situacionales que, se puede asumir, reflejan la influencia de aspectos naturales de jerarquía cultural, emocional o psicológica al momento de la elección. Por la forma de preguntar, la elección del jefe o jefa de familia se efectuó de manera natural, por lo que refleja influencias inadvertidas sobre el informante adecuado.

De nuevo, el problema es que esa clasificación, con base en la jerarquía, no existe en la base de datos original, pero sí las variables para tabularla. La tabla I presenta cuatro variables. Las primeras tres están incluidas en la base de datos original de la ENIGH, y la tercera variable se requeriría para efectuar el estudio con base en la "nueva" clasificación de familia. La variable *folio* es el número único identificador de todos los miembros de un grupo familiar. La primera familia la integran tres miembros; tres, la segunda familia, y cinco la tercera, y así sucesivamente. La variable *parentesco* manifiesta la relación que cada integrante de una familia guarda con el jefe o jefa identificado.

La variable *sexo* tiene los códigos 1 para hombre y 2 para mujer. La variable *Rela* (relación familiar) es la que se crea para rotular a cada uno de los integrantes de las familias como miembros de los nuevos tipos de familia con base en la jerarquía establecida por el informante apropiado: PC-Padre con Compañera, con código 1; MS-Madre sin Compañero, con código 2; PS-Padre sin Compañera, con código 3; y MC-Madre con Compañero, con código 4.

Efectuar las tareas de rotulación manualmente es impráctico en la ENIGH (se trata de 91738 personas en 22595 familias) y muy inexacto, por lo que es necesario utilizar herramientas de programación fácilmente disponibles. Enseguida se presenta un ejemplo de código de programación en Fox Pro para Windows, y una explicación breve de cada una de las instrucciones; claro, otros lenguajes de programación, como el Visual Basic o sus versiones más avanzadas, logran el mismo objetivo.

Tabla I. Cuatro tipos de variables.

Folio	Parentesco	Sexo	Rela
20040110010	Jefa(e)	Femenino	*MS*
20040110010	Hijo(a)	Femenino	*MS*
20040110010	Hijo(a)	Masculino	*MS*
20040110011	Jefa(e)	Masculino	*PC*
20040110011	Esposo(a)	Femenino	*PC*
20040110011	Hijo(a)	Femenino	*PC*
20040110020	Jefa(e)	Femenino	*MS*
20040110020	Hijo(a)	Femenino	*MS*
20040110020	Hijo(a)	Masculino	*MS*
20040110020	Hijo(a)	Masculino	*MS*
20040110020	Nieta(o)	Femenino	*MS*
20040110030	Jefa(e)	Masculino	*PS*
20040110030	Hermana(o)	Femenino	*PS*
20040110030	Sobrino(a)	Masculino	*PS*
20040110030	Sobrino(a)	Masculino	*PS*
20040110040	Jefa(e)	Femenino	*MS*
20040110040	Hijo(a)	Femenino	*MS*

```
go top                          Puntero enviado al primer registro
do while .t.                    Ejecute la tarea mientras haya registros
c=recno()                       Mando del contador secuencial de Fox Pro
e=folio                         Se almacena el valor de folio en "e"
f=c+1                           El valor de "c" más 1 para efectuar la comparación
s=sexo                          El valor de sexo: 1 hombre, 2 mujer
goto f                          Se repite el bucle para efectuar la comparación
g=folio                         Se almacena el valor de folio (familia) en "e"
if e=g                          Si todavía el integrante es parte de la misma familia
goto c                          Si sí es cierto, entonces efectúe la siguiente
                                comparación
do case                         Como no es cierto, se efectúa un bucle de caso
case s=1                        Si sexo es masculino
replace Aux with "PC"           La variable Aux= "PC"
case s=2                        Si sexo es femenino
replace Aux with "MC"           La variable Aux= "MC"
endcase                         Finaliza el bucle de caso
skip                            Se salta un registro
en la base de datos
endif                           Se termina el bucle "if"
replace Aux with "999"          En caso de no existir registro
                                de sexo
enddo                           Termina el bucle principal
```

Este programa o algoritmo presenta únicamente la solución para rotular a los padres que tienen una compañera y a las madres que tienen un compañero, y se presenta únicamente con propósitos de ejemplificación. Sin embargo, el algoritmo se empleó en la rotulación real en varias etapas. Presentaremos enseguida algunos de los resultados de manera muy escueta con base en la clasificación novedosa de las familias.

El porcentaje más alto de familias en esta nueva clasificación es el de las familias clásicas, en las que el hombre es jefe de familia (familias PC: 68.7%). Sin embargo, el porcentaje de las familias en el que la madre tiene la jerarquía principal, y no tiene compañero (MS), es bastante alto (20.5).

La tabla II presenta la asistencia a la escuela para los hijos e hijas entre los 5 y los 23 años en cada tipo de relación familiar. Las familias más eficientes, en cuanto a este indicador, son las de padre con compañera (PC) y las de madre con compañero (MC), mientras que las menos eficientes son las de padre sin compañera (PS) y las de madre sin compañero y MS.

La tabla III presenta el monto en pesos que invierte cada tipo de familia para gastos educativos. Las familias madre con compañero (MC) son las que más dinero invierten; las familias que menos invierten son las de madre sin compañero (MS).

La tabla IV presenta el promedio de horas que trabaja el jefe o jefa por tipo de relación familiar. El padre en las relaciones PC trabaja más horas, mientras que la jefa en las relaciones MC trabaja menos horas.

La tabla V presenta el promedio de horas trabajadas por el compañero(a) en los dos tipos de relación posibles: en las relaciones PC la compañera trabaja menos horas que las que trabaja el compañero en las relaciones MC.

La tabla VI presenta la ubicación de cada tipo de familia por tipo de rubro de gasto. El número 1 refleja el primer lugar

Tabla II. Frecuencia y porcentaje de los tipos de familia.

	Frecuencia	Porcentaje
Relación de jefe(a) de familia		
Padre jefe de familia con compañera (PC)	15532	68.7
Padre jefe de familia sin compañera (PS)	1898	8.4
Madre jefa de familia con compañero (MC)	535	2.4
Madre jefa de familia sin compañero (MS)	4630	20.5
	22595	100%

Tabla III. Promedio de instrucción de jefe o jefa de familia.

	(PC) (N=15532)	(MC) (N=535)	(PS) (N=1898)	(MS) (N=4630)
Promedio de Instrucción	3.47	3.20	3.30	2.73

Tabla IV. Asistencia a la escuela de hijos e hijas por tipo de relación, sexo de los hijos y edad (5-23 años).

		Jefe con compañera (PC)		Jefa con compañero (MC)		Jefe solo (PS)		Jefa sola (MS)	
Asistencia a la Escuela		**FREC.**	**%**	**FREC.**	**%**	**FREC.**	**%**	**FREC.**	**%**
Sí	Hijos	9140	77.5	249	68.0	122	51.9	1518	69.3
	Hijas	8940	79.5	252	71.4	114	55.1	1449	71.1
No	Hijos	2658	22.5	117	32.0	113	48.1	671	30.7
	Hijas	2309	20.5	101	28.6	93	44.9	590	28.9

(la cantidad más alta), mientras que el número 4 refleja el cuarto lugar (la cantidad más baja invertida).

Estas posibilidades de análisis estadístico, y otras que se encuentran en el documento completo, no hubieran sido factibles si no se hubiera logrado la rotulación mediante programación sobre la base de datos original. Los estudiantes de maestría y doctorado, así como los investigadores y profesores, deben explorar las posibilidades analíticas del análisis secundario de datos. No debe desaprovecharse la enorme cantidad de información de encuestas nacionales e internacionales ya disponibles en los bancos de datos. La mayoría son gratuitas, y representan enormes y emocionantes posibilidades analíticas para el investigador y el estudiante de posgrado.

Tabla V. Promedio de gastos educativos por tipo de familia.

	PC (8872)	MC (N=280)	PS (N=443)	MS (N=2141)
Artículos y servicios de educación	3932	4975	3576	3105

Tabla VI. Promedio de horas trabajadas por jefe(a) por relación.

Relación	N	Hrs./Sem.
PC	15532	45.97
MC	535	22.72
PS	1898	37.86
MS	4630	23.91

Tabla VII. Promedio de horas trabajadas por compañero(a) por relación.

	N	Hrs./Sem
PC	5759	38.56
MC	425	51.25

Tabla VIII. Lugar ocupado por cada familia.

	PC	MC	PS	MS
GASTOS				
Alquiler	2	1	4	3
Gasto total	2	1	3	4
Vestido y calzado	2	1	3	4
Esparcimiento	3	2	1	4
Pagos al banco	3	2	1	4
Enseres/ muebles	2	1	3	4
INVERSIONES				
Cuentas/ ahorro	2	3	1	4
Cuidados médicos	3	2	1	4
RIESGOS				
Tabaco	2	4	1	3