Badler, Clara
Alsina, Sara
Beltrán, Celina
Fracchia, Fernando
Puigsubirá, Cristina
Vitelleschi, Ma. Susana
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística.

EL MONITOREO COMO TÉCNICA DE DETECCIÓN DE LA FALTA DE INFORMACIÓN EN ENCUESTAS SOCIO-ECONÓMICAS

1. INTRODUCCIÓN

El interés por incrementar la calidad de la información provenientes de encuestas y registros ha ido creciendo, teniendo en cuenta que los datos que se obtienen, generalmente, presentan problemas que afectan los resultados. La evaluación de los mismos excede a los objetivos del análisis que el usuario se ha propuesto.

Por lo tanto, las bases de datos resultantes, presentan información faltante y/o confusa que en muchos casos puede ser ignorada porque se supone que ellas han sido tratadas previamente o que dicho tipo de información es natural, sobre todo en variables sensible como el ingreso.

En este trabajo se presenta la técnica del monitoreo de variables aplicadas a ondas de la Encuesta Permanente de Hogares (EPH) correspondiente al Aglomerado Gran Rosario, como instrumento imprescindible para la detección de información faltante y/o confusa.

2. MATERIAL

El material usado en este trabajo es heterogéneo dado los objetivos del mismo.

Bases correspondientes a la EPH, ondas: octubre 1997, mayo, agosto y octubre 1998 y mayo y octubre 1999, para el Aglomerado Gran Rosario. El archivo de trabajo es el de personas de la base usuaria ampliada.

Bloques: "desocupados" y "ocupados".

Cuestionarios y manuales de procedimientos: correspondientes a la EPH son utilizados como referentes. En relación a la codificación de la categoría "no sabe/no contesta" y las observaciones faltantes por diseño, en ellos se especifica: "IMPORTANTE: los códigos 9, 99, 999, 9999 corresponden en todos los casos a la categoría (no sabe/no responde). El código "0" identifica los casos a los cuales no le corresponde la secuencia analizada".

Se utilizan los softwares SAS 8.1 y Microsoft Excel 2000.

3. METODOLOGÍA

Se adopta como definición de monitoreo a: “El seguimiento y aplicación de pruebas de consistencia a información obtenida de encuestas resultantes de las diferentes etapas que constituyen las mismas: diseño, trabajo de campo, codificación, tabulación, etc.”.

Además, la información confusa y/o faltante es un problema multicausal, pero frecuentemente, es debido a la diversidad de definiciones y códigos que se plantean para la identificación de las mismas.

El monitoreo se realiza a través de las siguientes etapas:

I. Importación de la base de datos: se verifica que el proceso de importación capte la información sin modificar los datos de la base, dado que el archivo que contiene la información que se desea estudiar, normalmente tiene un formato diferente al que trabaja el software utilizado para analizar los datos.

II. Listado de la estructura de la base: se detalla la ubicación, el tipo de variable y la extensión de cada campo para todas las variables que contiene la base, con el objeto de comprobar la presencia y ubicación en la base.

III. Obtención de distribuciones de frecuencias para las variable: se controla si los valores observados son contemplados según la definición, en los casos que sea posible.

IV. Identificación de la información faltante: se analiza la presencia de códigos especiales o de casilleros en blanco, teniendo en cuenta la forma de identificación de la misma. Se debe contar con información respecto a la forma de identificación de la información faltante en la base de datos y si es posible diferenciarla de la falta por diseño.

V. Seguimiento de casos que presentan información faltante: se identifican las pérdidas parciales y totales.

VI. Realización de pruebas de consistencia: se intenta diferenciar la información perdida, la faltante por diseño y la confusa, a través del estudio de la cantidad de individuos que contestan cada pregunta.

VII. Describir patrones de asociación entre diferentes variables: se utiliza un modelo log-lineal obtenido a partir de la información presentada en las tablas de contingencia construidas al aplicar las pruebas de consistencia.

3. RESULTADOS

Se utilizan los softwares. SAS y Microsoft Excel para importar las bases de datos de la EPH, que se encuentran en archivo Dbase. En este proceso de importación no se presentan dificultades ya que los datos son captados sin tener modificaciones.

Luego se confecciona un listado de todas las variables mediante el PROC CONTENTS de SAS, verificando la presencia y características de las mismas a través de los manuales.

Para verificar las características (total de observaciones, categorías y códigos incluidos) de los valores observados en cada una de las variables, se realizan las distribuciones de frecuencias mediante el PROC FREQ de SAS.

Los manuales de la EPH son muy claros en cuanto a la definición de las categorías "No sabe / no responde" y la categoría "faltante por diseño". La definición textual que se encuentra es: "IMPORTANTE: los códigos 9, 99, 999, 9999 corresponden en todos los casos a la categoría (no sabe/no responde). El código (0) identifica los casos a los cuales no le corresponde la secuencia analizada".

Aunque no tendrían que surgir diferencias, ya que la definición comprende a todas las respuestas posibles, en las distribuciones de frecuencias se identifican algunas situaciones que no son previstas según las definiciones encontradas en los manuales, como por ejemplo la existencia de blancos.

La etapa siguiente del monitoreo lleva a realizar un seguimiento de los casos que presentan códigos previstos para la información faltante (9, 99, 999, 9999), para determinar si se trata de pérdidas totales o parciales, pero debido a la presencia de estos casos con información confusa, se decide realizar el seguimiento de los casos que presentan cero (para verificar si se trata de información faltante por diseño como define el manual) y de los casos que presentan blancos (para determinar a que categoría pertenecen).

El seguimiento de los casos demuestra que se presentan las siguientes situaciones para las dos categorías en estudio:

- "no sabe/no responde": blanco; cero; código 12.
- "no corresponde": blanco; cero.

En los casos en que los dos tipos de información faltante se presenta como blanco, el software que se utiliza para el análisis de la información, tomará la misma como perdida, pero para los casos de cero ó código 9 y 12, este último sólo presente en las variables que recategorizan los ingresos económicos, lo tomará como información real.

En las variables que representan el ingreso total familiar (ITF) y el ingreso per cápita familiar (IPCF), este hecho se ve acentuado, dado que las distribuciones de frecuencias de estas variables no muestran la presencia ni de 9, ni de blancos. Lo que hace pensar en la no presencia de información faltante. Pero dado que en este tipo de variables sensibles la presencia de información faltante es muy frecuente y en algunos casos de gran tamaño, se decide realizar un análisis detallado de la información disponible.

La característica más llamativa, es la presencia de una cantidad importante de ceros, lo que hace pensar en que todos los miembros del hogar tienen ingreso nulo, pero al realizar un seguimiento de estos casos, se comprueba que se mezclan los casos de hogares donde todos los miembros declaran tener ingreso nulo, con hogares donde algún miembro del hogar declara tener ingreso, pero no declara el monto y por definición de las variables ITF e IPCF, las mismas toman el valor nulo para todos los miembros del hogar.

Se realizan algunas pruebas de consistencia, para verificar la presencia de otras situaciones particulares, encontrando algunos inconvenientes con las cantidades de respondientes en algunas preguntas, como por ejemplo:

- Las unidades que en la secuencia del cuestionario registran las siguientes categorías de respuesta: P05= 3, 4, 5, u 8; ó P06= 1; ó P06= 2 y P07= 2, deberían pasar a la pregunta P17. Sin embargo se verifica que éstas contestan todas las preguntas desde la 12 a la 17.
- P40B se debería aplicar sólo a los que respondieron "9" en P40, sin embargo se observa que se reúne la información de la pregunta también para los que respondieron P40=5.

Una vez que se verifican cuales son las unidades que presentan información perdida, se realiza un conteo de la misma variable por variable, que permite sacar algunas características .

Para las variables exclusivas del bloque de desocupados la mayor frecuencia de pérdidas las presentan las variables relacionadas al tiempo transcurrido desde su ocupación anterior y con las características del establecimiento donde trabajaba.

Para las variables comunes a los bloques de desocupados y de ocupados, se observa que las que presentan mayor proporción de información faltante son las relacionadas con el ingreso, pero se encuentra también falta de información en variables relacionadas a inmigración y características de ocupación.

Se establece la proporción de variables con falta de información en las 19 variables con pérdidas comunes a los dos bloques para las distintas ondas analizadas (Tabla 1) y su clasificación según el porcentaje de unidades afectadas en cada una (Tabla 2).

Tabla 1. Número de variables con falta de información según onda y bloque

Bloques	Ondas											
	Oct-97		May–98		Agos-98		Oct-98		May-99		Oct-99	
	Nº	%	Nº	%	Nº	%	Nº	%	Nº	%	Nº	%
Desocupados	7	37	6	32	3	16	5	26	5	26	6	32
Ocupados	12	63	9	47	11	58	13	68	11	58	10	53

Tabla 2. Cantidad de variables con falta de información según porcentaje de unidades afectadas, bloque y onda

Porcentaje de variables	Ondas											
	Oct–97		May–98		Agos-98		Oct-98		May-99		Oct-99	
	Des.	Ocu.	Des.	Ocu.	Des.	Ocu.	Des.	Ocu.	Des.	Ocu.	Des.	Ocu.
(0.01 – 1]	5	5	3	2	0	3	4	6	1	5	0	5
(1 – 2]	1	2	2	2	1	3	0	2	3	1	4	0
(2 – 5]	0	0	1	0	2	0	0	0	0	0	1	0
(5 y +)	1	5	0	5	0	5	1	5	1	5	1	5

Variables relacionadas al ingreso

Dadas las características especiales encontradas para estas variables, se presenta el número de unidades con el valor "cero" y las que se comprueban como faltantes en el monitoreo para los bloques de desocupados y de ocupados en cada onda (Tabla 3).

Tabla 3. Cantidad de unidades con valor cero verificadas como información faltante según variables de ingresos, bloque y ondas

VARIABLE	ONDA Oct-97		ONDA May-98		ONDA Ago-98		ONDA Oct-98		ONDA May-99		ONDA Oct-99	
	Desoc. n=206	Ocup. N=1299	Desoc. n=214	Ocup. n=1260	Desoc. n=162	Ocup. n=1177	Desoc. n=170	Ocup. n=1033	Desoc. n=159	Ocup. n=873	Desoc. n=176	Ocup. n=856
P47T	163 (1)	124 (89)	163 (1)	99 (69)	120 (0)	102 (79)	129 (0)	108 (92)	115 (3)	109 (98)	140 (2)	107 (90)
ITF–IPCF	33 (12)	126 (122)	22 (7)	96 (94)	21 (7)	124 (123)	40 (15)	126 (125)	22 (19)	144 (144)	37 (18)	124 (124)

Con el objeto de evaluar la relación entre la presencia de ceros en la variable ITF, el estado ocupacional y las distintas ondas, se ajusta un modelo loglineal.

Para obtener independencia entre las frecuencias de las celdas de la tabla de contingencia a analizar, se selecciona un representante de cada hogar, excluyendo los hogares que son habitados solo por inactivos. Para los hogares donde habitan uno o más desocupados se selecciona un desocupado y para los hogares donde no habitan desocupados se selecciona un ocupado.

Las variables consideradas en este análisis son:

- Estado: con las categorías Ocupados (1) y Desocupados (2) coincidentes con las categorías de la base.
- ITF (Ingreso Total Familiar): considerada dicotómica con las categorías ITF=0 (0) e ITF>0 (1)

- Onda: se consideran tres ondas correspondientes a octubre 97, agosto 98 y octubre 99.

Al relacionar estas variables se obtiene la siguiente tabla de contingencia:

Tabla 4. **Cantidad de personas clasificadas según estado, onda e ingreso total familiar**

Estado	Ondas						
	Oct 97		Agos 98		Oct 99		
	ITF = 0	ITF > 0	ITF = 0	ITF > 0	ITF = 0	ITF > 0	Total
Ocupado	64	652	69	605	64	405	1859
Desocupado	29	144	18	116	30	117	454
Total	93	796	87	721	94	522	2313

El modelo seleccionado mediante el procedimiento "backward" contiene todas las asociaciones de a dos variables: (ITF*Onda, Onda*Estado, ITF*Estado), siendo el valor de la estadística G^2=1.26 (p=0,5326; g.l.=2). En este modelo, la asociación entre dos variables no varía con las categorías de la tercer variable, lo que implica que la asociación existente entre ITF y Estado es la misma en las tres ondas (Tabla 5).

Tabla 5. Parámetros estimados del modelo

Parámetro	Estimación	Error estándar
ITF	-0,9265	0,0366
ESTADO	0,6054	0,0367
ITF*ESTADO	-0,1297	0,0367
ONDA	0,0956	0,0492
	-0,0438	0,0509
ITF*ONDA	-0,0765	0,0453
	-0,0519	0,0462
ONDA*ESTADO	0,00608	0,0361
	0,1050	0,0381

La interpretación del modelo se realiza a través de las razones de odds asociadas (Tabla 6).

Tabla 6. Razones de odds estimadas del modelo e intervalos de confianza del 95%

	Estimación de las razones de odds	Intervalo de confianza del 95%
ITF*ESTADO	0.595	(0.4465;0.7936)
ITF*ONDA		
Onda 97 vs. 98	0.9520	(0.6980;1.2980)
Onda 97 vs. 99	0.6638	(0.4876;0.9036)
ONDA*ESTADO		
Onda 97 vs. 98	0.8205	(0.6394;1.053)
Onda 98 vs. 99	1.5406	(1.0470;2.266)

En las tres ondas, los ingresos nulos se presentan principalmente en aquellos hogares donde habitan desocupados. Asimismo, en los hogares encuestados en la onda de octubre de 1999 se observan ingresos familiares nulos más frecuentemente que en las ondas de agosto de 1998 y octubre de 1997.

4. DISCUSIÓN

La técnica de Monitoreo debe ser aplicada previamente a la descripción y análisis de bases de datos provenientes de encuestas y registros.

Conjuntamente con la evaluación de la calidad de la información esta técnica permite detectar la información faltante y/o confusa que no siempre es directamente visualizable. Para su implementación se requiere la consideración simultánea de manuales de procedimientos, técnicas estadísticas y elementos computacionales.

La información faltante y/o confusa puede surgir en cada una de las etapas necesarias para la realización de encuestas y registros y obedecen a problemas multicausales, pero al detectarse pueden aplicarse tratamientos para mejorar la calidad de los datos a analizar.

Bibliografía

Agresti, A..(1990).”Categorical Data Analysis”. J.Wiley & Sons.NY.

Aster, R. and Seidman, R..(1997).“Professional SAS Programming Secrets”. McGraw-Hill.

Badler,C.;Alsina,S.;Wojdyla,D.;Fracchia,F..”Interpretación y Procesamiento de Bases de Datos con Información Faltante a través de Softwares Estadísticos”. Presentado al XXVIII Coloquio Argentino de Estadística de la Soc. Arg. de Estadística. Posadas, 2000.

INDEC.(1998).Encuesta Permanente de Hogares. “Explicación para el uso de la base usuaria ampliada en GBA”.

Little, R. J. and D. B. Rubin. (1987). “Statistical Analysis with Missing Data”. John Wiley & Sons. New York.

Platek, R. and C. Särndal. (2001). “Can a Statistician Deliver?. Journal of Official Statistics, vol. 17, Nº 1, pp. 1-20.

SAS®.(1990).Procedures Guide: Version 6, Third Edition, Cary, NC: SAS Institute Inc.

SPSS® (1997).Base 9.0, Applications Guide, SPSS Inc.