REVISION AUTOMATICA DE DATOS DE CENSOS Y ENCUESTAS MEDIANTE EL USO DE COMPUTADORES MEDIANOS Y PEQUEÑOS

Julio Ortúzar
(CELADE)

AUTOMATIC EDITING OF CENSUS AND SURVEY DATA USING A SMALL OR MEDIUM SIZE COMPUTER

SUMMARY

This article examines problems of editing population surveys and census data. An outline is given of the principal steps involved in editing: error detection; correction of inconsistencies and the documentation of the errors detected and corrected. The CONCOR (CONsistency and CORrection) system, designed by CELADE for the purpose of carrying out these steps on small and medium computers, is described indicating the posibilities which the present system offers and possible future improvements.

I. INTRODUCCION

Durante muchos años, la revisión de datos provenientes de censos y encuestas y su corrección se han constituido en uno de los grandes cuellos de botella en la elaboración de la información estadística, especialmente en países en vías de desarrollo, donde las oficinas de estadística no siempre cuentan con especialistas en procesamiento de datos con suficiente experiencia. Más aún, en la mayoría de estos países se dispone sólo de pequeños computadores, lo que también contribuye a dificultar esa tarea.

En este artículo se presenta un procedimiento relativamente sencillo, utilizado por el Centro Latinoamericano de Demografía (CELADE)

en el procesamiento de encuestas y censos de población y vivienda, en países que han solicitado la asesoría del Centro. Tal procedimiento se apoya en gran medida en el sistema CONCOR (CONsistency and CORrection) desarrollado por CELADE con el objeto de facilitar la tarea de revisión de la información básica.

Definición del problema

A través de las diferentes etapas por las que pasan los datos en el curso de su elaboración, se producen errores que afectan la información y tienden a falsear la realidad del universo en estudio. El objetivo fundamental de la revisión es detectar dichos errores, corregirlos y documentarlos, de tal forma que los usuarios tengan conocimiento de la calidad de los datos con que trabajarán.

En esa tarea existen dos etapas claramente diferenciadas:

1. La que tiene por objeto verificar la representación del universo estudiado en el archivo de datos, esto es, que no existan omisiones ni duplicaciones de unidades de registro.

2. La que se destina a verificar si la información que está registrada es consistente. Se dice "consistente" y no "correcta" puesto que en ocasiones un dato o un código puede estar mal registrado y ser coherente con los demás, lo que impide que el error sea detectado.

Generalmente, la primera de estas etapas se realiza en forma paralela al traspaso de la información desde el documento fuente a tarjetas, cinta magnética o disco, mientras la segunda se realiza una vez finalizada la creación del archivo maestro.

II. DETECCION DE ERRORES

Existen, pues, dos niveles distintos de errores, que exigen procedimientos diferentes para ser detectados:

Representatividad del universo estudiado

La primera preocupación en la tarea de revisión debe ser la de obtener un archivo de datos que represente fielmente el universo investigado. Dejando de lado los problemas de cartografía y de muestreo, siempre cabe la posibilidad de que el archivo final no refleje fielmente los cuestio-

narios recogidos en el campo, en particular cuando se manejan grandes volúmenes de información, como es el caso de los censos.

Omisión o duplicación de casos

Debido a la manipulación repetida del material recogido, es relativamente fácil que se produzcan omisiones o duplicaciones, ya sea de casos aislados o, lo que es peor, de unidades completas de empadronamiento o de muestreo. Así, por ejemplo, una de estas unidades podría no llegar nunca a la sección de digitación (entrada de datos), o bien podría serle enviada dos veces. Análogamente, una "cassette" o una caja de tarjetas podría omitirse o ingresarse más de una vez al archivo maestro.

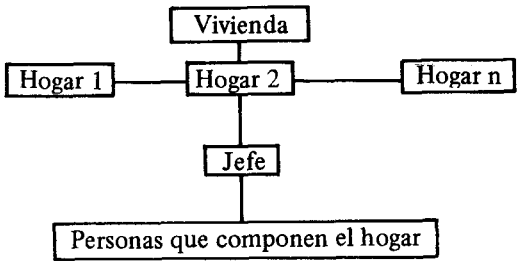
Para prevenir este problema en el momento oportuno es necesario, además de una buena organización, un control estricto de la información que se ingresa al archivo maestro. Este control se puede ejercer mediante la comparación de totales de casos ingresados al archivo maestro a nivel de pequeñas unidades geográficas, administrativas, o de otra naturaleza, con totales provenientes de otras fuentes de información. Asimismo, una vez que cada unidad es incorporada al archivo maestro, el sistema de manejo de datos utilizado debería rechazar cualquier otra que tenga la misma identificación. El sistema de manejo de datos, deberá a su vez emitir informes de elaboración que permitan analizar las diferencias observadas, unidades aceptadas y rechazadas, etc.

Omisión o duplicación de registros

A menudo la unidad de estudio se compone de varios registros, siendo a veces, la estructura de la información muy simple, y en otras, bastante compleja. Sin embargo, en cualquiera de los casos es posible que se omita o duplique uno o más de estos registros, error que evidentemente no será detectado con el procedimiento indicado. En la figura 1 se puede observar una estructura bastante simple de un caso que puede ser asimilado a un censo de población y vivienda. Si los registros están bien diseñados, cada unidad debe indicar el número de registros que tiene la unidad que le sigue. Así, en el caso presentado en la figura 1, el registro de vivienda debe contener el número de hogares que la habita; a su vez, cada registro de hogar debe indicar el número de personas que lo forma.

Teniendo esto en consideración, resulta fácil asegurarse de que la estructura de cada caso esté correcta.

Figura 1



Consistencia de la información registrada

Una vez que se obtiene un archivo completo de datos, que contenga todos los casos que corresponda y sólo esos, y que cada uno de ellos esté estructuralmente correcto, corresponde preocuparse de los errores deslizados en la información registrada. Para la detección de este tipo de errores, conviene prescindir del concepto de "registro", y considerar simplemente un caso como un conjunto de elementos de información o variables. Bajo este contexto, existen dos métodos para detectar errores en la información.

Verificación de rango

Puesto que siempre se sabe cuáles son los posibles códigos de una variable discreta, y siempre se podrá establecer un límite inferior y otro superior para las variables continuas, se puede detectar un error en la información registrada verificando si el código registrado es válido (de acuerdo al conjunto de códigos previamente definido para cada variable en particular), o si es un valor que está dentro del intervalo establecido.

Reglas de consistencia entre variables

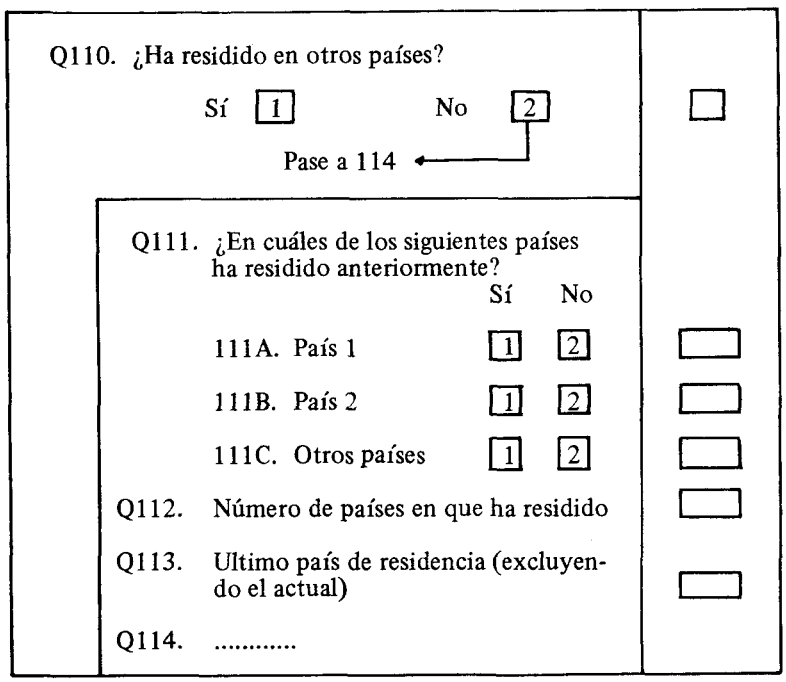
Dado que una variable puede contener un error, y sin embargo pasar satisfactoriamente la verificación de rango, es posible aprovechar las relaciones que existen entre las variables para establecer reglas de consistencia entre ellas e identificar errores no detectados anteriormente. Estas relaciones pueden ser clasificadas en tres tipos:

a) *Relaciones dadas por el flujo del cuestionario*

Por lo general, tanto en encuestas como en censos hay preguntas o variables que se investigan sólo para una parte del universo estudiado.

Las variables que definen a este subconjunto son, en consecuencia, las que determinan que otras variables tengan o no información. Si se verifica el flujo del cuestionario se puede determinar la omisión de información a nivel de variable, o a la inversa, la presencia de información cuando ella no debe existir. Para analizar esta situación, se puede tomar el ejemplo que se presenta en la figura 2.

Figura 2



En este caso, las siguientes reglas permitirán detectar un error en el flujo del cuestionario:

R_1 : (Q110 = 2) ⟷ (Q111A = blanco)

R_2 : (Q111A = blanco) ⟷ (Q111B = blanco)

R_3 : (Q111B = blanco) ⟷ (Q111C = blanco)

R_4 : (Q111C = blanco) ⟷ (Q112 = blanco)

R_5 : (Q112 = blanco) ⟷ (Q113 = blanco)

b) *Relaciones lógicas entre variables*

Con frecuencia existe entre las variables investigadas este tipo de relación, lo que permite detectar otra clase de errores. Continuando con el ejemplo de la figura 2, se establecen las siguientes relaciones:

R_6 : (Q110 = 1) $\longleftrightarrow$ (Q111A=1) $\cup$ (Q111B = 1) $\cup$ (Q111C = 1)

R_7 : (Q113 = 1) $\longleftrightarrow$ (Q111A = 1)

R_8 : (Q113 = 2) $\longleftrightarrow$ (Q111B = 1)

R_9 : (Q113 = 3) $\longleftrightarrow$ (Q111C = 1)

c) *Relaciones aritméticas*

Para las variables de tipo cuantitativo se pueden establecer relaciones aritméticas que permiten acotar el límite de variación admisible. Considérense dos variables: *EDAD* y *AÑOS DE INSTRUCCION.* Para fines censales, la primera puede oscilar entre 0 y 99, y la segunda entre 0 y 20. La diferencia entre la primera y la segunda variable no puede ser inferior a 5, 6 ó más, conforme a las normas del país en estudio.

Cabe destacar que el hecho de considerar la unidad de estudio como un conjunto de variables, sin tener en cuenta el registro al cual pertenece, permitirá detectar errores que muchas veces pasarían desapercibidos si se realizara la revisión a nivel de registro. Así, por ejemplo, en los censos de población de la región latinoamericana ha sido frecuente que se planee la detección de errores a nivel de individuos, sin considerar las relaciones que existen entre personas que pertenecen a un mismo hogar. Ello ha suscitado problemas posteriores, cuando se emprenden estudios de familias y resultan, por ejemplo, hogares en los que el jefe es menor que su propio hijo, o muestra un estado civil distinto al de su cónyuge.

III. CORRECCION DE ERRORES O ELIMINACION DE INCONSISTENCIAS

Se indicó inicialmente que, tanto la verificación de la representatividad de los datos en relación al universo estudiado como de su estructura, representan una etapa previa a la verificación de consistencia entre variables. La tarea de eliminación de los errores detectados en la primera etapa es bastante más simple de lo que puede llegar a ser la labor misma de identificarlos. En efecto, detectado un error de este tipo, será necesario efectuar una de las siguientes operaciones:

> *Intercalaeión:* si hubo omisión de casos habrá que intercalarlos en el subarchivo en el orden que corresponda.

Eliminación: si hubo duplicación de casos, habrá que eliminarlos del subarchivo.

La corrección de errores en la segunda etapa de la revisión es mucho más compleja, y en ella se puede llegar a invertir una apreciable cantidad de tiempo. Para esta labor, existen dos posibilidades:

1. *Corrección manual*

Por este medio, el computador se limita sólo a detectar el error, quedando la responsabilidad de la corrección en el personal encargado de la encuesta. El computador proporciona la información que permite:

a) Localizar e identificar el cuestionario o documento con errores;

b) Individualizar la variable que presenta divergencia en sus respectivos códigos o valores.

Mediante el análisis del documento original se determinan los cambios que hay que efectuar en la información almacenada en el archivo maestro, y por medio de algún procedimiento se efectúan los cambios necesarios. Esto es aparentemente muy simple. Sin embargo, involucra una cantidad de problemas:

i) Al analizar un documento que contiene un error o incoherencia, se puede encontrar que la información almacenada en el archivo maestro difiere de la registrada en el documento fuente, en cuyo caso el error se produjo en alguna etapa intermedia entre la recolección de los datos y su almacenamiento en el archivo maestro. La solución será simplemente el reemplazo de la información errónea por la que aparece en la fuente.

Si la información almacenada es idéntica a la del documento fuente, el error proviene de la etapa de recolección. En ese caso, la situación puede ser más compleja. A veces la respuesta o el código puede ser deducido a través del análisis de otras preguntas. En otras, sin embargo, será imposible hacerlo sin volver al informante. En tales casos, lo más que se podrá hacer es acotar el posible valor (en el caso de variables cuantitativas) o el código correspondiente a "ignorado". Es en esta situación donde el procedimiento de corrección manual resulta limitado.

ii) Al introducir cambios en un cuestionario, hay que someterlo nuevamente al programa de consistencia, puesto que, al cambiar las condiciones iniciales, nuevos errores pueden ser detectados, lo que contribuye a transformar la corrección en un proceso iterativo.

iii) El proceso manual contiene las deficiencias propias de tales procedimientos: es lento; se pueden introducir nuevos errores en vez de

corregir los antiguos; los criterios de corrección no se aplican en forma sistemática; los procedimientos de asignación son muy limitados y pueden introducir equivocaciones; la documentación de errores corregidos es deficiente.

De los problemas enunciados se puede deducir que el procedimiento de corrección de errores en forma manual puede ser apropiado o aconsejable para el caso de una encuesta pequeña, pero difícilmente podrá serlo para el caso de un censo. En una encuesta:

a) Los datos recolectados son generalmente de mejor calidad, ya que los encuestadores cuentan normalmente con capacitación especial. Esto debe implicar que parte de los errores son introducidos con posterioridad a la recolección de datos, y por tanto su solución se encuentra en el cuestionario mismo.

b) Se utiliza, en todas las etapas de elaboración de datos, un menor número de personas y con una mejor capacitación, siendo, por tanto, relativamente pequeña la cantidad de errores introducidos en esta etapa. Esta situación, unida a un volumen no muy grande de datos manejados, reduce el tiempo necesario para la tarea de limpieza.

c) Se investiga, por lo general, un gran número de variables y por consiguiente el programa de limpieza es largo y complejo. En tal caso, es muy probable que la programación de esta operación lleve mucho más tiempo que la corrección manual, y pueda retrasar el procesamiento de los datos.

2. *Corrección automática*

Esta es la fase más compleja y controvertida de la revisión, y sobre ella existen opiniones antagónicas. Hay quienes son partidarios de que se impute en forma automática la información faltante o registrada erróneamente, y quienes por el contrario, prefieren conservar una categoría de información "no declarada" para cada variable. El problema es que, aunque no se haga ninguna asignación durante la etapa de elaboración de los datos, el usuario de la información muchas veces tendrá que hacerlo posteriormente. Aun en el caso que se decida no considerar en el análisis aquellas unidades que están en categoría desconocida, se está, de hecho, admitiendo que la distribución de estos valores es proporcional al resto del universo para el cual se posee información.

Reconociendo que la corrección automática es un problema delicado, en este artículo sólo se describe una metodología, sin entrar en la polémica de quienes la defienden o la rechazan.

En la corrección automática existen dos problemas básicos:

a) *Identificación de las variables que deben ser corregidas.*

Cuando se comprueba que una variable falla en la verificación de rango, dicha variable pasa de inmediato a la lista de aquéllas que deberán ser corregidas. Sin embargo, cuando una regla de consistencia involucra dos o más variables, y se observa una divergencia entre ellas, hay que decidir cuáles de ellas deben ser corregidas, de modo de eliminar la inconsistencia. I.P. Felligi & D. Holt desarrollaron una metodología para la identificación de lo que ellos llaman "conjunto mínimo de campos para imputación". Esta metodología parte de la idea básica de representar en una matriz las variables y las reglas de consistencia que se establecen para ellas, pero difiere de otras en la forma de seleccionar el conjunto de variables que se va a corregir.

A fin de facilitar la explicación del método, utilizemos nuevamente el ejemplo de la figura 2, representando en las columnas de la matriz cada una de las variables, y en las filas las reglas de consistencia. Para cada regla se identifica con "1" las variables involucradas y con "0" a las demás. En esta forma se construye una matriz de 9 x 6 (nueve reglas y seis variables):

Figura 3

	Q110	Q111A	Q111B	Q111C	Q112	Q113	E
R_1 :	1	1	0	0	0	0	
R_2 :	0	1	1	0	0	0	
R_3: :	0	0	1	1	0	0	
R_4 :	0	0	0	1	1	0	
R_5 :	0	0	0	0	1	1	
R_6 :	1	1	1	1	0	0	
R_7 :	0	1	0	0	0	1	
R_8 :	0	0	1	0	0	1	
R_9 :	0	0	0	1	0	1	
F							

Se asocia a la matriz un vector columna (E), que contendrá el resultado de la evaluación de cada una de las reglas de consistencia. Un "1" significa que las variables de un registro específico satisfacen la regla de consistencia, y un "0" que no la satisfacen. Se asocia también un vector fila (F) que contendrá la frecuencia con que cada variable estuvo involucrada en una regla no satisfecha, dividida por la frecuencia de participación de la variable en el total de reglas establecidas. Finalmente, se aplican las siguientes normas:

i) Se elige como variable a ser corregida aquélla que muestra un coeficiente más alto en el vector *F.*

ii) Entre variables que tienen igual coeficiente, se elige aquélla que participe en más reglas de consistencia.

iii) Si aún hay más de una posibilidad, se elige una variable en forma arbitraria.

iv) Elegida la variable a corregirse, se supone que todas las reglas de consistencia en que ella participa serán satisfechas (se cambian los respectivos ceros por unos en el vector *E).*

v) El procedimiento continúa con la reevaluación de los vectores y repitiendo los pasos *(i)* a *(iv)* hasta que todas las reglas queden satisfechas.

El algoritmo depende fundamentalmente de las reglas de consistencia establecidas, y no existiendo un programa aplicable a computadores pequeños que, *i)* identifique las reglas de consistencia contradictorias, y *ii)* determine las reglas de consistencia implícitas o derivadas de las explicitadas, se hace necesario analizar con extremo cuidado las relaciones entre las diferentes variables en estudio. Este problema se puede apreciar claramente en la continuación del ejemplo analizado: supóngase que las variables de la figura 2 lleven los siguientes códigos:

Q110 = 1; Q111A = blanco; Q111B = 2; Q112 = 2 y Q113 = 2.

En este caso, deliberadamente, se han introducido errores, de tal modo que fallen las dos relaciones en que interviene la variable Q110. Ello implicará que Q110 sea seleccionada como una de las variables a ser corregida, en circunstancias que está obviamente correcta. Sin embargo, la situación se plantea en forma distinta si se agregan las siguientes reglas de consistencia, derivadas de las primeras cinco:

R_{10}: (Q110 = 1) ⟷ (Q111B ≠ blanco)

R_{11}: (Q110 = 1) ⟷ (Q111C ≠ blanco)

R12: (Q110 = 1) ⟷ (Q112 ≠ blanco)

R_{13}: (Q110 = 1) ⟷ (Q113 ≠ blanco)

Ahora, aplicando el algoritmo especificado, resulta que fallan las reglas 1, 2, 6 y 8. Haciendo la evaluación de los coeficientes, resulta seleccionada la variable Q111A. En segunda instancia, solamente falla la regla 8 en que intervienen las variables Q111B y Q113. Puesto que ambas participan en el mismo número de relaciones, se elige una en forma arbitraria.

En la figura 4 se esquematiza cada uno de los pases del procedimiento de selección.

b) *Corrección de las variables*

En el procedimiento de corrección existen dos situaciones, y para cada una de ellas hay un método de corrección diferente:

i) Cuando se sabe *a priori* que existe un solo valor posible que satisface el conjunto de relaciones o reglas de consistencia establecidas. En este caso el procedimiento de corrección es "deductivo", debiendo hacerse un análisis de los códigos o características de las variables que se relacionan con la que se desea corregir y, basado en ellas se determina el código a imputar. Puesto que puede haber entre las variables relacionadas algunas que aún no hayan sido corregidas, es necesario descartar el procedimiento deductivo para dichas variables, lo que contribuye a complicar el método.

Sin embargo, el procedimiento puede ser fácilmente reemplazado por una tabla con tantas entradas como variables se relacionen con la que se desea corregir y en cada una de sus celdas el código que corresponda a la variable que se está corrigiendo al conectar las variables relacionadas de las líneas con las variables relacionadas de las columnas, por un operador "AND". A fin de aclarar esta idea, en la figura 5 se presenta una tabla para corregir la variable Q110, presentada en el ejemplo anterior. El "*" como código de las variables de preferencia, corresponde a una "marca" que identifica a la variable que fue seleccionada para ser corregida, y aún no lo ha sido. La "X" como código que se va a imputar, identifica a una combinación imposible. Puesto que en esta etapa debe haberse identificado ya todas las variables en conflicto, tales combinaciones imposibles

Figura 4

R	Q110	Q111A	Q111B	Q111C	Q112	Q113	E'	E"	E"'
R_1	1	1	0	0	0	0	0	1	1
R_2	0	1	1	0	0	0	0	1	1
R_3	0	0	1	1	0	0	1	1	1
R_4	0	0	0	1	1	0	1	1	1
R_5	0	0	0	0	1	1	1	1	1
R_6	1	1	1	1	0	0	0	0	0
R_7	0	1	0	0	0	1	1	1	1
R_8	0	0	1	0	0	1	0	1	1
R_9	0	0	0	1	0	1	1	1	1
R_{10}	1	0	1	0	0	0	1	1	1
R_{11}	1	0	0	1	0	0	1	1	1
R_{12}	1	0	0	0	1	0	1	1	1
R_{13}	1	0	0	0	0	1	1	1	1
F'	2/6	3/4	3/5	1/5	0	1/5			
F"	0	0	1/5	0	0	1/5			
F"'	0	0	0	0	0	0			

no pueden ocurrir. Supóngase que M_{ij} es el código de la línea "i" y columna "j" de la tabla con la que se quiere imputar a la variable Q110. Entonces, definir Q110 = M_{35}, equivale a la siguiente expresión: (Q111A = 1 ∪ Q111A = 2) ∩ Q111B = * ∩ (Q111C = 1 ∪ Q111C = 2) ∩ Q112 = * ∩ (Q113 = 1 ∪ Q113 = 2 ∪ Q113 = 3) ⟶ Q110 = 1 (Celda correspondiente a la línea 3 de la columna 5), en la que las variables Q111B y Q112 deben aún ser corregidas. Nótese que cada una de las celdas de la tabla da origen a una expresión similar. Localizar la celda de la tabla es mucho más simple, rápido en término de tiempo de computador, y ocupa menos memoria.

FIGURA 5

					VALORES DE Q112						
					blanco (b)		1 ó 2 ó 3		asterisco (*)		
					AND						
					VALORES DE Q113						
					b	*	2 ó 3	*	1 ó 2 ó 3	b	*
	Q111A		Q111B	Q111C							
1	1 ó 2		1 ó 2	1 ó 2	x	x	1	1	1	x	1
2				*	x	x	1	1	1	x	1
3	1 ó 2		*	1 ó 2	x	x	1	1	1	x	1
4				*	x	x	1	1	1	x	1
5	*		1 ó 2	1 ó 2	x	x	1	1	1	x	1
6				*	x	x	1	1	1	x	1
7	*		*	1 ó 2	x	x	1	1	1	x	1
8		AND		b	2	2	x	x	x	2	2
9				*	2	2	1	1	1	2	2
10	*		b	b	2	2	x	x	x	2	2
11				*	2	2	x	x	x	2	2
12	b		b	b	2	2	x	x	x	2	2
13				*	2	2	x	x	x	2	2
14	b		*	b	2	2	x	x	x	2	2
15				*	2	2	x	x	x	2	2
					1	2	3	4	5	6	7

ii) Cuando se sabe *a priori* que hay más de un código que satisface las relaciones de consistencia establecidas, se busca un procedimiento de entrada que permita relacionar las diversas variables que intervienen en las reglas de consistencia más aquellas que muestran algún grado de conexión con la que se va a corregir.

Supóngase, por ejemplo, que se desea revisar la variable "hijos nacidos vivos", la que se encuentra correlacionada por lo menos con otras tres variables: "estado civil"; "edad"; y "nivel de instrucción". Cualquiera que sea el método de asignación, estas variables deberán ser tomadas en cuenta. Sin embargo, lo más probable es que sólo la variable "edad" esté relacionada con "hijos nacidos vivos", mediante una relación de consistencia.

El procedimiento de corrección que corrientemente se usa es el del "hot deck", que consiste en ubicar entre los registros ya procesados, y que cumplieron satisfactoriamente las reglas de consistencia establecidas, el último que muestre características similares a aquéllas que no están en conflicto en el registro que se está procesando. Este procedimiento no es aplicable en computadores pequeños, y por tanto se hace necesario una simplificación:

a) Eliminación de variables que no estén relacionadas con la que está siendo corregida.

b) Estratificación de variables relacionadas, según la capacidad del computador disponible. Por ejemplo: para imputar "hijos nacidos vivos", la variable estado civil se podría estratificar en *i)* solteras; *ii)* casadas y unidas; y *iii)* viudas, divorciadas y separadas. Luego se reemplaza la búsqueda de un registro determinado por una tabla de tantas entradas cuantas sean las variables seleccionadas. Esta tabla se define inicialmente con valores provenientes de otras fuentes, y es actualizada por cada registro que no presenta inconsistencia en la variable que se está corrigiendo.

La figura 6 ilustra una tabla de 28 líneas y 21 columnas, que permitiría imputar el número de hijos nacidos vivos en función de las tres variables que más fuertemente se correlacionan con ese dato. El "*" nuevamente indica que la variable a la cual se refiere debe ser aún corregida. En este caso, la actualización de una celda de la matriz deberá hacerse con el mismo valor las celdas marcadas con asterisco.

Así, una mujer de 18 años, casada y sin instrucción, que tiene 1 hijo nacido vivo, deberá reemplazar por 01 las siguientes celdas: *i)* (8,4) que corresponde a casada sin instrucción y 18 años; *ii)* (8, 21) que corresponde a casada sin instrucción; *iii)* (14, 4) que corresponde a casada y 18 años; *iv)* (14,21) que corresponde a casada; *v)* (22,4) que corresponde a 18 años sin instrucción; *vi)* (22,21) correspondiente a sin instrucción;

Figura 6

Estado civil e instrucción	E D A D									
	15	16	17	...29	30-34	35-39	40-44	45-49	50 y+	*
	1	2	3	15	16	17	18	19	20	21
Solteras										
1. Sin instrucción										
2. Primaria incomp.										
3. Primaria comp.										
4. Secundaria incomp.										
5. Secundaria comp.										
6. Universit. y Sup.										
7 *										
Casadas-Unidas										
8. Sin instrucción										
9. Primaria incomp.										
10. Primaria comp.										
11. Secundaria incomp.										
12. Secundaria comp.										
13. Universit. y Sup.										
14. *										
Viudas, Separadas Divorciadas										
15. Sin instrucción										
16. Primaria incomp.										
17. Primaria comp.										
18. Secundaria incomp.										
19. Secundaria comp.										
20. Universit y Sup.										
21. *										
22. Sin instrucción										
23. Primaria incomp.										
24. Primaria compl.										
25. Secundaria Incomp.										
26. Secundaria comp.										
27. Universit y Sup.										
28. *										

vii) (28,4) correspondiente a 18 años y *viii)* (28,21) que correspondera simplemente a la última mujer mayor de 14 años.

Algunos comentarios a este método

Mientras más variables que se correlacionan con la que se quiere asignar se tomen en cuenta, y mientras menos agrupadas estén sus respectivas categorías, mejor reproducirán la distribución de frecuencia condicionada de los casos sin asignar.

Es probable que en el ejemplo dado, la fecundidad de las mujeres casadas sea diferente a la de las mujeres en unión libre, o que la fecundidad de dos mujeres de 30 y de 34 años sea diferente.

En resumen, lo óptimo sería encontrar un registro gemelo del que se está corrigiendo, y a base de él hacer la asignación. El grado de detalle a que se puede llegar dependerá básicamente de la capacidad del computador disponible.

El método presupone que los errores se encuentran distribuidos al azar, y por tanto que los datos no han sido ordenados según un concepto especial, diferente al de su ordenamiento natural.

IV. DOCUMENTACION DE ERRORES DETECTADOS Y CORREGIDOS

Esta es una etapa posterior a la limpieza de los datos, aunque no menos importante que las dos anteriores. Ella es la que permite juzgar, hasta cierto punto, si los procedimientos aplicados han sido adecuados, además de cuantificar las modificaciones hechas en la información básica.

La documentación de los errores detectados y corregidos es de dos clases:

1. Información global que permite obtener una primera impresión del proceso de limpieza. Tal información puede ser, por ejemplo, la distribución de frecuencia marginal de los casos asignados, tanto en valores absolutos como en porcentaje en relación al total.

2. Información detallada en un archivo que contenga:

   a) Cada uno de los registros de los casos con error, una vez corregidos.

   b) Identificación de las variables que fueron corregidas, con sus respectivos códigos o valores antes de la asignación.

Con esta información se pueden obtener cruces para comparar las frecuencias condicionadas de ambos conjuntos de datos (los asignados y los no asignados) y si es necesario, inspeccionar en forma minuciosa la decisión adoptada en determinados casos. Puesto que esta tarea puede consumir muchos recursos del computador, normalmente se efectúa como una segunda etapa, limitándose la primera a detectar y corregir los errores y producir el archivo ya especificado en alguna memoria de respaldo, como disco o cinta magnética.

V. UN SISTEMA PARA LA REVISION DE DATOS ESTADISTICOS: CONCOR

De lo expuesto anteriormente pueden deducirse las condiciones que un sistema de revisión debería reunir para abarcar todas las fases de la limpieza de datos estadísticos. Para formarse una idea general de la potencialidad del sistema CONCOR, se presentará el esquema de las facilidades que él ofrece. Para ello se clasifica como *AI* aquellas facilidades que se encuentran "actualmente incorporadas"; como *CP* las que se incorporarán a "corto plazo", y como *MP* las que se incluirán a "mediano plazo" en el sistema CONCOR.

Puesto que CONCOR debía ser instalado en computadores pequeños, fue escrito en lenguaje Assembler IBM 360/370, con el objeto de ocupar relativamente poca memoria (mínimo 48K bytes) y tener un máximo de eficiencia en la ejecución.

Actualmente el "International Statistical Program Center del US Bureau of the Census" está desarrollando en COBOL, en conjunto con el CELADE, una versión de CONCOR, que incorporará nuevas funciones para la revisión de datos.

El sistema está dividido en 5 programas centrales (ver figura 7) que pueden ser ejecutados en fases consecutivas, a opción del usuario:

MONITOR: programa de comunicación entre el usuario y el sistema, a través del cual el usuario especifica qué funciones del sistema desea utilizar.

COMPILER: analizador sintáctico e intérprete del lenguaje CONCOR, que se encarga además de generar archivos en disco (opcionalmente en cinta) que contienen *a)* tablas y constantes derivadas del programa del usuario, y *b)* diccionario del archivo de entrada y mensajes de error.

El lenguaje contempla 6 diferentes tipos de comandos:

1. Recodificación y conversión de datos.

2. Evaluación de expresiones aritméticas.

3. Funciones de verificación.

4. Control de flujo del proceso.

5. Asignación dinámica (Hot deck).

6. Generación de registros de salida.

EXECUTOR: Programa que ejecuta las operaciones especificadas por el usuario en su propio programa, escrito en lenguaje CONCOR y traducidas por COMPILER.

LISTERR: Programa que se encarga de producir toda la documentación y estadísticas de los errores detectados por EXECUTOR.

CORRECTOR: Programa que se utiliza sólo cuando se desea efectuar corrección manual de los datos. Su función es la de efectuar modificaciones en el archivo de datos básicos especificadas por el usuario a través de comandos *ad hoc*.

Esquema de las facilidades de un sistema generalizado de revisión

	Facilidades	*Clasificación*
1.	*Para detectar y corregir errores de representatividad:*	
	a) Comparación de total de casos a nivel de estratos o unidades geográfico-administrativas con totales de otras fuentes.....	MP
	b) Chequeo de estructura.....	CP
	c) Eliminación, intercalación y reemplazo de casos completos o registros aislados.....	AI
2.	*Para el chequeo de consistencia:*	
	a) Verificación de rangos.....	AI
	b) Consistencia de relaciones lógicas entre variables.....	AI
	c) Relaciones aritméticas entre variables.....	AI
	d) Flujo del cuestionario.....	AI

e) Identificación de reglas de consistencia contradictorias..... MP

f) Derivación de reglas implícitas..... MP

3. *Para la corrección de errores en forma directa:*

a) Acceso al cuestionario, registro y variable de un archivo..... AI

4. *Para la corrección automática:*

a) Identificación de variables con error..... CP

b) Manejo de tablas de varias entradas..... AI

c) Método de asignación automática *i.e.* "Hot deck"..... AI

5. *Documentación de errores:*

a) Mensajes identificando cuestionario y variables con error..... AI

b) Control por el usuario del nivel de detalle en la documentación de errores..... AI

c) Archivo de casos con error identificando variables en conflicto: valores antes de la corrección y valores después de la corrección..... AI

Experiencias con el Sistema CONCOR

El sistema CONCOR ha sido aplicado con éxito en numerosas encuestas, muestras de censos de población y vivienda, y en un censo completo.

Son ejemplos de su aplicación:

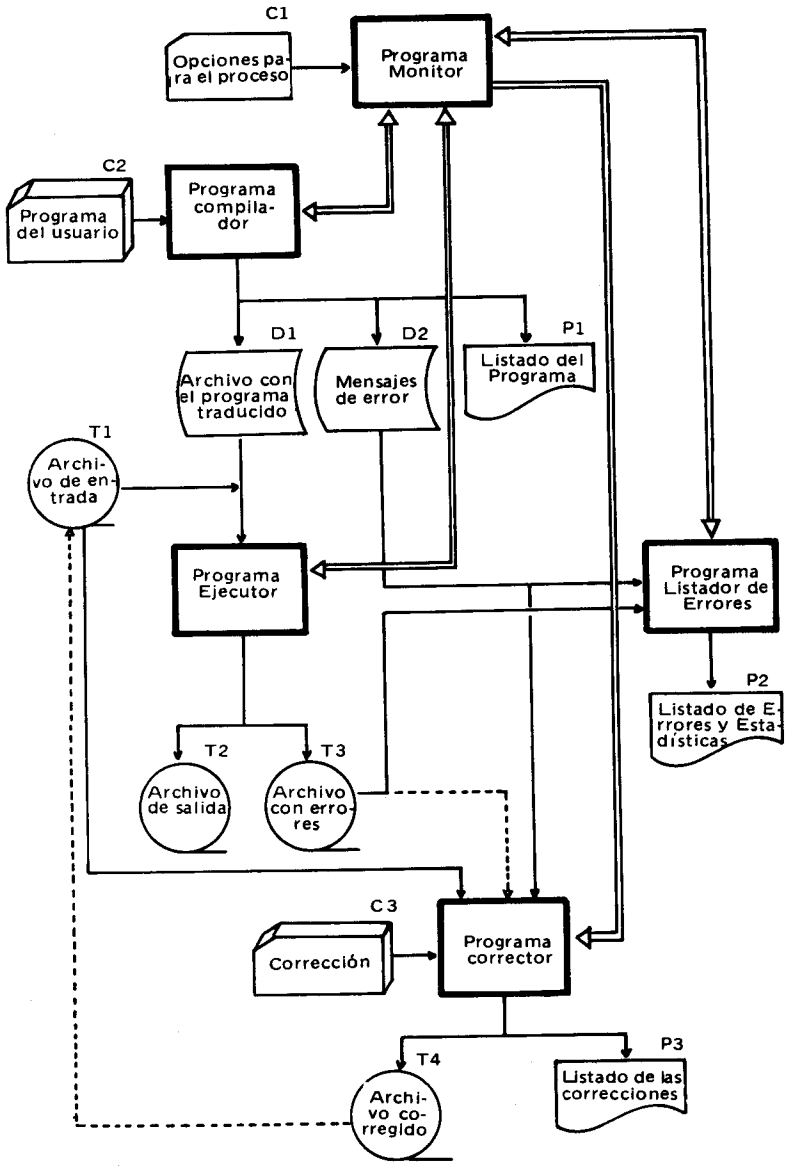
a) Encuestas Demográficas de Bolivia y Perú;

b) Encuestas Nacionales de Fecundidad de República Dominicana, Panamá, Costa Rica y Colombia;

c) Censos de Haití y Uruguay;

d) Muestras de censos de casi todos los países latinoamericanos.

FIGURA 7


C1
Opciones para el proceso
Programa Monitor
C2
Programa del usuario
Programa compilador
D1
Archivo con el programa traducido
D2
Mensajes de error
P1
Listado del Programa
T1
Archivo de entrada
Programa Ejecutor
Programa Listador de Errores
P2
Listado de Errores y Estadísticas
T2
Archivo de salida
T3
Archivo con errores
C3
Corrección
Programa corrector
T4
Archivo corregido
P3
Listado de las correcciones

Adicionalmente a su objetivo básico, CONCOR ha sido usado eficientemente como "inter-fase" entre el archivo de datos básicos y algún sistema de tabulación. Usualmente, estos sistemas no aceptan como entrada un archivo jerárquico, por lo que se ha usado esta capacidad de CONCOR para transformar el archivo y a la vez crear índices y variables recodificados que posteriormente servirán de entrada para el sistema de tabulación.

Refiriéndose al problema representado por la limpieza de la información, o "data editing", se expresó Sir Maurice Kendall, Director de la "World Fertility Survey":

"For decades the most underdeveloped phase of the processing of social surveys has been that of data editing, or cleaning. Typically it has been undertaken using either unit-record equipment or specially written survey-specific computer programs. For a survey as complicated as ours the utilization of such techniques would involve anywhere from 8 to 24 months of work by a dedicated staff. But in the last few years more giant steps have been taken to develop general purpose editing and imputation programs. Thus, CANEDIT (formerly known as GEISHA) has been developed by Statistics Canada, and CONCOR by CELADE (The Latin American Demographic Centre).

"After a thorough evaluation of both systems,and after considering the feasibility of developing our own general purpose editing system, the WFS adopted to use CONCOR. We are working closely with CELADE in continuously adapting and improving the various components of CONCOR. Partly because of our adoption of CONCOR, the interval between the end of field work and the start of tabulations has been reduced from a range of 10 to 16 months in the case of early WFS surveys to an average now of about 4 to 6 months. In addition, since CONCOR is a generalized system, once installed in a country, it may be used for cleaning other surveys, and even censuses. In this way the WFS is co-operating with CELADE in the diffusion of modern technology for the processing of surveys".

BIBLIOGRAFIA

(1) Felligi, I.P. and Holt, D.A., *A sistematic approach to automatic edit and imputation.*

(2) Graves, R.B., *CANEDIT a Generalized Edit and Imputation System in a Data Base Environment.*

(3) Kendall, M., *The Analysis of World Fertility Survey Data.*