I N F O R M E   P R E L I M I N A R

DESCRIPCION DEL PROGRAMA DE VALIDACION Y COHERENCIA DE

DATOS PARA LA MUESTRA CENSAL DE NICARAGUA ;

Julio Ortúzar

CELADE

19 de mayo de 1972

DESCRIPCION DEL PROGRAMA DE VALIDACION Y CORRECCION DE DATOS PARA LA MUESTRA CENSAL DE NICARAGUA

I. INTRODUCCION:

Antes de empezar con la descripción del programa mismo y de los criterios que se utilizaron para corregir errores, es necesario aclarar las bases sobre las que este se diseñó, a la vez que las restricciones que contiene.

Dado el escaso tiempo con que se contaba para tener el programa funcionando, nos dedicamos a analizar solamente los datos referentes a población, dejando para una segunda etapa el estudio de la información de vivienda. Aún así, la depuración de los datos de población puede hacerse en forma más profunda, camino por el cual pensamos seguir avanzando. Es por ello que esta descripción, además de representar lo que el actual programa analiza, pretende completar lo que pensamos debe ser en una etapa más avanzada.

A fin de tener una estimación global de la calidad de los datos, y a fin de saber con qué profundidad debíamos analizar determinadas variables para obtener un dato más depurado, sacamos una distribución de frecuencias marginales de cada una de las variables del registro de población. Esto a su vez nos indicó por qué variables no nos debíamos preocupar pues no presentaban códigos inválidos o un porcentaje de casos no declarados muy apreciable. Por otra parte, ella nos da también una estimación del porcentaje de casos que requerirán de una asignación de código, en base al cual es posible decidir o programar la metodología más adecuada para esta asignación.

II. DESCRIPCION DEL PROGRAMA

El programa consta de 4 etapas claramente diferenciadas, y ellas son las que pasamos a describir a continuación, en el mismo orden en que son ejecutadas:

1. Verificación de la validez de códigos de cada una de las variables estudiadas:

En esta etapa se detecta variable a variable todos aquellos valores que no estén contemplados como códigos válidos, inclusive la ausencia de información cuando ésta deba venir.

Cada una de las variables estudiadas tiene una rutina, la cual se encarga de analizar el error cuando este es detectado por el programa principal. En algunos casos, la rutina, investigando otras variables o por algún procedimiento aleatorio, asigna un determinado código diferente de aquel que signifique "ignorado". En otros, simplemente asigna el código equivalente a "ignorado".

Veamos a continuación la metodología seguida para la asignación de cada una de las variables:

a) Lugar de empadronamiento:

Para asignar un código ya sea a Departamento, Municipio o zona urbano-rural, se tiene en cuenta el código del último individuo que se procesó. Hay que tener cuidado al emplear este criterio, puesto que si la información ha sido ordenada ("sorteada") según estos conceptos, no sería adecuado. Para que él sea válido, los datos deberívenir tal como salen de perforación. Es decir, los individuos de una misma boleta, uno a continuación de otro, y ojalá, las boletas de un mismo sector, una a continuación de otra. Para clarificar más las ideas, vamos a ver las posibilidades con un ejemplo.

Código de Dep.
Antes de Ordenar

01	03	06
02	23	09
02	03	08
02	20	07
04	05	⋮
		16

Código de Dep.
Después de Ordenar

01	04	10
02	05	⋮
02	06	16
02	07	23
03	08	
03	09	

El código ilícito es 23, ya que solo existen 16 departamentos.

Se ve en forma clara, que si asignamos antes de ordenar la información, asignaríamos un código 03 que es el que está inmediatamente antes. En cambio, si lo hacemos en igual forma, una vez la información esté ordenada, asignaríamos 16. Como se ve, lo más probable es que el individuo con código inválido pertenezca al departamento 03.

Se podría ser aún más estrictos en la asignación con este criterio, pero a nuestro juicio, no se justifica debido al reducido número de errores que en general se producen con este tipo de datos.

b) Características Generales:

- Relación con el jefe del hogar: Para esta variable no hay ningún análisis en esta etapa. Simplemente se asigna el código que signifique "ignorado". Posteriormente, en otra etapa, se investigará si el "hogar" tiene jefe o no. Si este falta, se creará un jefe.

- Sexo: Para asignar esta variable, existen dos criterios. Cual de ellos se use depende fundamentalmente de la edad del individuo en cuestión. Si es de 15 o más años de edad, se analizan los campos referentes a fecundidad. Si ellos vienen en blanco se asigna sexo "masculino". En caso contrario, "femenino". El límite de edad dependerá por supuesto de cada país) Si es menor de 15 años, se asignará en forma aleatoria, tomando el sexo contrario al último individuo que se procesó.

Como se ve, el primer criterio exige que la fecundidad se le haya perforado solo a la población femenina.

Una crítica que se podría hacer en contra de este criterio, es el utilizar datos que aún no han sido "validados", como serían edad y fecundidad. Nuestro descargo se basa en la pequeñísima cantidad de datos que traerían estas dos o tres variables con error. Si pequeño es el número de casos con error en sexo, menor lo será aún el número que trae error simultáneo en las 2 variables, debido a lo cual lo despreciamos.

- Edad: Para la asignación de esta variable, se divide la población en 3 grupos de acuerdo a las características que vengan perforadas en la tarjeta:

El primer grupo queda limitado a edades fluctuando entre 0 y 5 años, y se define por aquella parte de la población que solo trae perforada información referente a características generales. El resto (culturales, económicas y fecundidad) viene en blanco. Para asignar una edad a este grupo, simplemente se toma aquella que mostró el último menor de 6 años que se procesó. El segundo grupo se define por aquella parte de la población que solo trae información en características generales y culturales, viniendo el resto en blanco. Las edades fluctuarán sólo entre 6 y 9 años. Para asignar una edad a este grupo, se sigue la misma metodología que para el grupo anterior, con la diferencia que se refiere a población entre 6 y 9 años. Además, una vez asignada la edad, se verifica que esta sea coherente con el nivel de instrucción que aparece codificado. Esto es, que la edad deberá ser mayor que el número de años de estudios equivalentes al nivel de instrucción más 4. Si esta relación no se cumpliese, se determinará la edad sumándole 4 al número de años de estudio. Si este valor resultara mayor que 9, se asignará nuevamente la edad con el primer criterio señalado., y se cambiará el nivel de instrucción a "no declarado".

Se podría pensar, para no complicar tanto la metodología, en asignar la edad solo en función del nivel de instrucción. Posiblemente esa filosofía sería adecuada en países donde el analfabetismo es muy pequeño y el ingreso a los colegios o escuelas se produce casi en forma uniforme a los 5 años de edad. Mas en el caso de Nicaragua, y la mayor parte de los países latinoamericanos, una gran parte de la población tiene cero años de estudio, razón por la cual pensamos que podría producir una sesgo importante.

El tercer grupo queda definido por el resto de la población y las edades fluctuarán entre 10 y más años.

Esta es la parte más delicada en la asignación para esta variable, tanto por el volumen de datos que en general hay que asignar, como por el tamaño del intervalo de edades posibles. Para llevar a cabo esta operación, tomamos en cuenta dos variables:

Relación con el jefe y tipo de actividad, las cuales las agrupamos en las siguientes categorías:

Relación con el Jefe	Tipo de Actividad
(1) Jefe (0)	(1) Tienen empleo y cesantes (1, 2, 3)
(2) Cónyuge (1)	(2) Buscaron trabajo por la. vez (4)
(3) Hijo (2)	(3) Jubilados (5)
(4) Padres o suegros (3)	(4) Estudiantes (6)
(5) Nieto (4)	(5) Cuidado del hogar (7)
(6) Empleada domést. (6)	(6) Otros e ignorados (8 y 9)
(7) Otros (5 y 7)	
(8) Ignorado (9)	

Creemos que la edad del último individuo procesado que mostró iguales características en estas dos variables es una buena metodología para asignar.

Si pensamos en las posibles edades del grupo (1) en Tipo de Actividad, vemos que son muchas. Probablemente entre 10 y 60 o 65 años. Mas si asociamos este grupo con el (3) en Relación con el Jefe, el intervalo de edades disminuye digamos entre 10 y 30 o 35 años. Si se quisiera ser aún más estricto con la asignación, se podría pensar en introducir una tercera variable que podría ser sexo.

Una vez asignada la edad mediante este procedimiento, queda por estudiarse si ella es coherente con el nivel de instrucción, o bien con el número de hijos tenidos si el individuo en cuestión es una mujer. Perfectamente se puede dar el caso que asignemos una edad de 10 años, y que el nivel de instrucción sea equivalente a 10 años de estudios o más, o bien que se trate de una mujer con dos hijos. Para ello, verificamos que se cumplan las siguientes relaciones:

Edad mayor o igual que años de estudio más 4

Edad mayor o igual que el número de hijos más 14
(en el caso de ser mujer)

Si alguna de ellas no se cumple, se asignará la edad mínima que cumpla con la(s) relacion(es).

Para el caso particular de esta variable, tomamos también como código inválido el equivalente a "ignorado".

- Estado Civil: En general, para esta variable se asigna simplemente el código equivalente a "ignorado" salvo cuando se trata de un jefe de hogar o cónyuge del jefe del hogar. En el primer caso, se ve si en el hogar existe un cónyuge, en cuyo caso se asigna el mismo estado civil que éste. Si no lo hay, se asigna ignorado. En el segundo caso, se asigna el mismo estado civil que trae codificado el jefe.

- Condición de Orfandad: Para esta variable solo se asigna el código equivalente a "ignorado" Podría pensarse en que si el individuo en cuestión es "hijo" del jefe del hogar, y existe otro hijo en el mismo hogar, asignar el código de este último ya que presumiblemente son hermanos "al menos de padre" (si el jefe es hombre). En todo caso, nosotros nos hemos limitado a asignar simplemente ignorado.

- Para todas aquellas variables que dicen relación con el lugar de nacimiento, residencia habitual y residencia anterior no se hacen asignaciones en esta etapa. Simplemente se detectan los códigos inválidos y son reemplazados por el correspondiente a "Ignorado."

c) Características culturales:

- Alfabetismo: Para asignar esta variable, se analiza el nivel de instrucción. Si éste es equivalente a segundo año de primaria o mayor, se asigna un código equivalente a "alfabeto". En cualquier otro caso, el código es reemplazado por el de "ignorado".

- Nivel de Instrucción: En esta etapa, solamente se asigna 4 códigos y ellos son los que equivalen a primaria, secundaria y superior con grado ignorado, e ignorado. Los tres primeros se asignan cuando el código de nivel está correcto (o supuestamente correcto), en cambio el código que se refiere al grado es ilícito. La asignación de "Ignorado" se efectúa cuando no es posible saber en qué nivel está el individuo, o bien hay ausencia de información.

- Asistencia escolar: En esta variable no se hace ninguna asignación. El código ilícito es reemplazado por el de "ignorado".

d) Características económicas:

Ocupación, Rama de Actividad, y Categoría ocupacional no son asignados en ninguna etapa, a diferencia de Tipo de Actividad que en una etapa posterior puede ser asignada. A esta altura, solo son reemplazados los códigos inválidos por el correspondiente de "ignorado".

e) Fecundidad:

- Hijos nacidos vivos: El análisis de esta variable consta de dos partes: una que es el reemplazo de los códigos inválidos por el de "Ignorado", y la otra que es un poco más compleja y que detallamos a continuación:

Dado el porcentaje apreciable de mujeres de las cuales se desconoce el número de hijos, y generalmente mayor en las edades jóvenes de la población femenina, fue que decidimos crear una nueva variable, la cual denominamos "Hijos nacidos vivos asignados" y experimentar alguna metodología de tal modo que pueda obtenerse un dato más "completo".

Si el dato viene completamente correcto, simplemente se copia en la nueva variable. Mas si no lo es, o trae el código equivalente a ignorado, se entra a asignar el número de hijos en función de tres variables que consideramos diferenciales para los distintos niveles de la fecundidad y que son: Estado civil, Nivel de instrucción y Edad. En base a estas 3 variables, buscamos la última mujer procesada que mostró iguales características en ellas y que traía la fecundidad correcta, y es asignada a la mujer en cuestión. En realidad, por razones de capacidad de memoria del computador, las categorías de las variables fueron agrupadas de la siguiente manera:

NIVEL DE INSTRUCCION	ESTADO CIVIL	EDAD
1) Primaria	1) Solteras	1) 15
2) Secundaria	2) Casadas y Unidas	2) 16
3) Superior	3) Viudas y Separadas	3) 17
4) Ignorada		4) 18
		5) 19
		6) 20 - 24
		7) 25 - 29
		8) 30 - 34
		9) 35 - 39
		10) 40 y más

Probablemente sería aconsejable abrir el nivel de instrucción a 7 categorías y el grupo de edad 20 - 24 por edades simples, a fin de obtener un dato más cercano a la realidad de la mujer en cuestión. En todo caso, es interesante comparar tabulaciones de fecundidad con ambas variables: "Hijos nacidos vivos" e "Hijos nacidos vivos asignados". A nuestro juicio es tan interesante este punto que pensamos seguir experimentando nuevas posibilidades. Experiencia que traduciremos en un informe posterior.

- Para el resto de las variables estudiadas en Fecundidad, hasta el momento nos hemos conformado con reemplazar los códigos inválidos por el correspondiente de "Ignorado".

2. Análisis de Coherencia entre Variables

a) Edad versus otras características: Como ya se explicó anteriormente, la edad debe estar de acuerdo a una cantidad de información que el individuo trae. Debido a ello es que el programa verifica que:

- Si es menor de 6 años, debe traer en blanco en toda la parte de la tarjeta que contiene información sobre características culturales, económicas y de fecundidad.

- Si es menor de 10 años, deberá traer en blanco la parte referente a características económicas y fecundidad.

En realidad, de encontrarse un error de este tipo, debería corregirse la edad, de tal modo de hacerla coherente con la información que el registro contiene. Desgraciadamente, las instrucciones para la codificación no se ciñeron en forma estricta a las normas dictadas tanto en la boleta como en el manual, apareciendo una apreciable cantidad de casos de menores de 6 años con información en características culturales, y de menores de 10 años con información en Tipo de Actividad. Después de un estudio de algunos de ellos, se concluyó que la edad estaba correcta, en cambio se había perforado información innecesaria. Por esta razón, las rutinas que analizan este tipo de error no corrigen la edad, pero sí acusan el error.

b) Edad versus Nivel de Instrucción: Para analizar posibles incoherencias entre estas dos variables, se traduce el Nivel en años de instrucción, el cual debe cumplir la relación Edad mayor que Años de Instrucción más 4. Si no se cumple, el nivel de instrucción es reemplazado por el más alto valor que se ajusta a la desigualdad.

c) Edad versus Hijos Nacidos vivos: Para la población femenina se verifica que se cumpla la relación Edad mayor que Número de Hijos más 13. En caso de no cumplirse esta relación, se asigna el código equivalente a ignorado en fecundidad, y se asigna un numero de hijos adecuado en la variable "hijos asignados".

d) Alfabetismo versus nivel de instrucción: Para todos los individuos mayores de 6 años que tengan segundo año de primaria o más en nivel de instrucción, se les verifica que no tengan un código distinto del de "Alfabetos" en la variable Alfabetismo. Si lo tuviesen, se asignará el código adecuado.

e) Tipo de Actividad versus Variables Económicas: Si un individuo aparece como "activo" según la variable Tipo de Actividad, se verificará que tenga datos en las restantes variables económicas. Por el contrario, se es "inactivo", se verificará que traiga en blanco los campos que se refieren a Ocupación, Rama de Actividad Económica y Categoría Ocupacional. Si no ocurre así, en el primer caso se asigna el código "Ignorado" a tipo de actividad, y en el segundo, se reemplaza el código por el de "Activo".

3. Conversión del Registro a Binario:

En esta etapa, todas las variables son convertidas a binario y grabadas en una cinta, la cual contendrá finalmente el archivo depurado en la forma ya descrita.

Análogamente, en otra cinta serán grabados todos los errores detectados, con datos suficientes que nos permitan identificar: las variables con error, el código asignado, y finalmente, la boleta en la cual podrá encontrarse al individuo en cuestión. Posteriormente saldrá una tabulación por variables con errores según códigos asignados. Esto nos permite tener una estimación de la cantidad de errores que hemos corregido, y de la frecuencia de códigos asignados.

4. Actualización de las Tablas de Distribución:

En este punto, y antes de pasar a un nuevo registro, se cambian o actualizan cada una de las tablas preparadas para asignación. Cabe señalar, sin embargo, que no necesariamente todas las tablas se actualizan, ya que, por ejemplo, la tabla de Fecundidad solo se actualiza si el registro que se ha procesado recientemente corresponde a una mujer de 15 años o más, y solamente si el dato sobre fecundidad venía correcto.

De esta forma, la asignación será totalmente dinámica, ya que los valores que se van asignando constantemente, van cambiando de acuerdo a los nuevos registros que se están procesando.

III. Conclusiones:

Creemos que es necesario destacar los siguientes puntos:

1. Para tener una buena asignación de lugar de empadronamiento, es absolutamente necesario que no se haga ningún ordenamiento mecánico o electrónico con las tarjetas. Si esto se hiciera, la metodología sobre la cual se basó "fracasaría en forma absoluta".

2. Si se pudiera procesar diariamente la producción de tarjetas perfo-verificadas de cada día con este programa, se podría controlar la calidad de la información entregada como un sub-producto.

3. Aconsejamos seguir experimentando en la muestra con pequeños cambios en la metodología de asignación para Edad y Fecundidad, especialmente en los valores iniciales de las tablas y agrupación de categorías.