DT12
Junio 1993

METODOLOGIA PARA EL TRATAMIENTO

DE LOS DATOS DEL FORMULARIO AMPLIADO[1]

Análisis Demográfico - Diseño Conceptual Censo 1991

Buenos Aires, mayo de 1993.

[1] Este documento fue elaborado por Gladys Massé bajo la coordinación de Alejandro Giusti. Los comentarios expresados son el resultado de reuniones de trabajo junto a Antonia Giangualani y Perla Zafrán.

Introducción

El presente documento tiene por objetivo explicar la metodología implementada para evaluar y consistir los datos obtenidos a partir de las variables precodificadas de1 formulario ampliado.

Se conoce que toda información censal y, en general, toda investigaci6n estadística que involucra un gran volumen de datos, presupone una alta probabilidad de que se presenten errores de diversa índole o magnitud. Ellos pueden originarse en diferentes etapas del procedimiento censal y estar asociados a causas diversas. También son variados los efectos que pueden ocasionar sobre la calidad de los datos censales. Por ese motivo, se considera de suma importancia proceder a la evaluación y consistencia de los datos del Censo Nacional de Población y Vivienda de 1991 como tarea previa a la tabulación de la información correspondiente al cuestionario ampliado, especificando posibles soluciones para cada caso concreto, con el fin de corroborar su calidad.

Los datos del formulario ampliado que también figuran en el básico -relación de parentesco, sexo, edad, asistencia, nivel y compleción del nivel-, así como las respuestas precodificadas de la variable lugar de nacimiento fueron consistidas en ocasión de procesarse el archivo básico y no se revisan en esta oportunidad. Las temáticas específicas del cuestionario ampliado que se consideran en este proceso son: material predominante de las paredes y cubierta exterior del techo de la vivienda; combustible utilizado por el hogar para cocinar; pensión o jubilación, residencia habitual, residencia habitual cinco años antes, cobertura de salud, alfabetismo, último grado aprobado, ítems relacionados con la condición de actividad, categoría ocupacional, jurisdicción (de los establecimientos públicos), tamaño (de los establecimientos privados), descuento jubilatorio, estado conyugal y fecundidad, todos ellos referidos a las personas.

Si bien la temática central para el estudio de las Necesidades Básicas Insatisfechas (NBI) 1991 refiere sólo a la elaboración de pautas para construir de manera eficiente el quinto indicador del índice -cuatro o más personas por miembro ocupado y, además, cuyo jefe tuviera baja educación-, dos son las razones que llevan a consistir todo el conjunto de variables. Por un lado, la interdependencia que existe entre ocupaci6n y educación respecto de otras variables investigadas en el censo -descuento jubilatorio, obra social, fecundidad-, por otro, la necesidad de aplicar el programa Concord[2] una sola vez para todas las variables del formulario ampliado. Este procedimiento, si bien trae aparejado la necesidad de mayor tiempo de dedicación, es la única vía que permite obtener de una sola vez informaci6n válida y coherente.

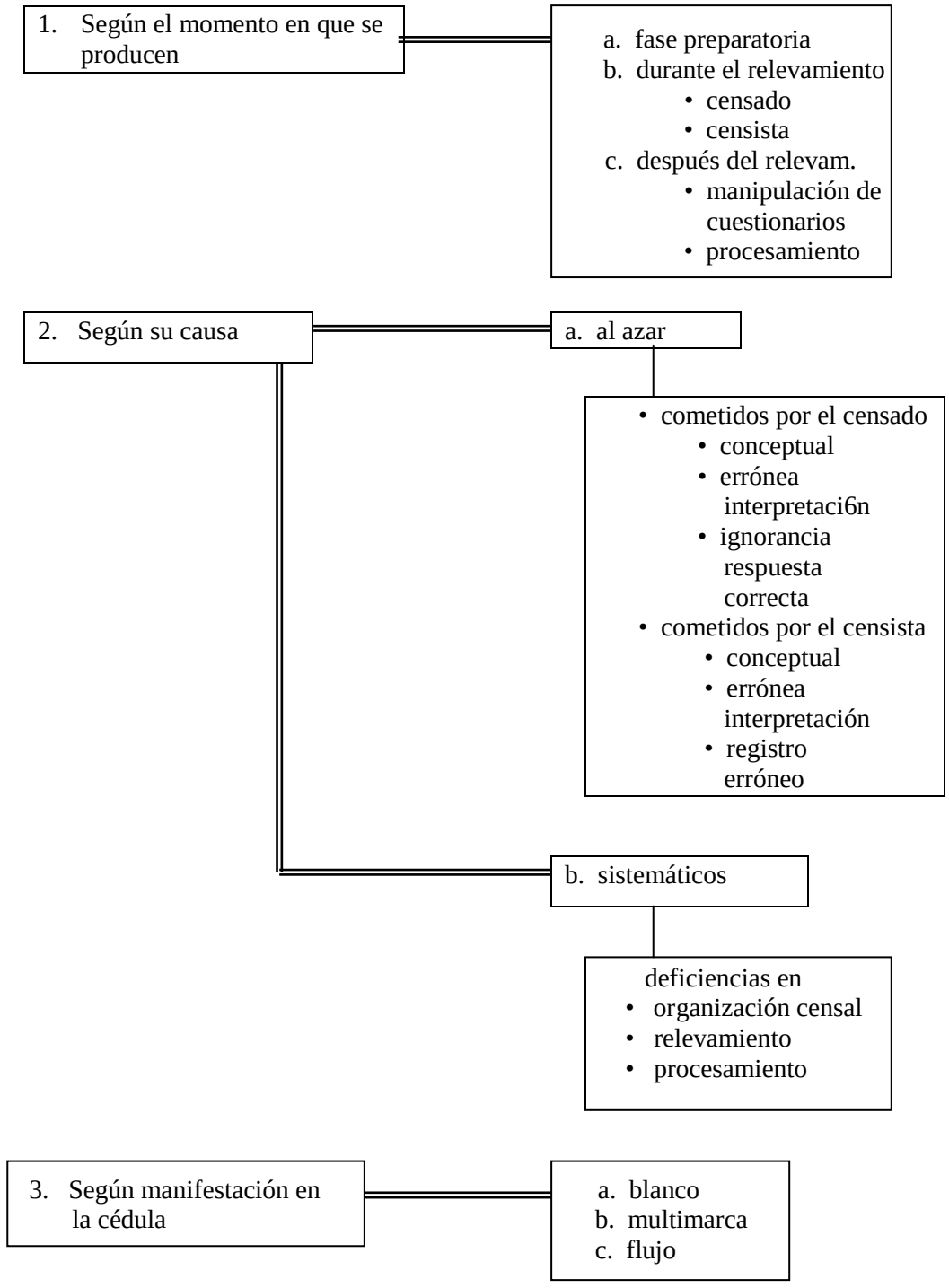
1. Los errores y sus características

Toda investigación demográfica se lleva a cabo a partir de información que puede estar afectada por diversos tipos de error. La magnitud de éstos también es variable y por ello es necesario indagar acerca de la cantidad y la calidad del error cometido, con el fin de decidir y luego efectuar una asignación de la forma más correcta posible y con bases científicas, disminuyendo la posibilidad de realizar un análisis sesgado de la realidad

En el siguiente gráfico puede observarse los diferentes tipos de errores que pueden introducirse en cualquiera de las diversas etapas del procedimiento censal:

[2] El Concord es un programa de asignación automática que asigna información a partir de pautas especificadas por el programador y establecidas en forma previa por el investigador.

Tipos de errores.



1. Según el momento específico en que éstos se producen:

a. Fase preparatoria del relevamiento:
Los errores pueden producirse por problemas en la preparación de la cartografía necesaria; errores en la elaboración de la muestra; insuficiente clarificación de los conceptos e instrucciones a utilizar; o una distorsión de dichos conceptos durante la etapa de capacitación de los censistas;

b. Durante el relevamiento los errores pueden ser cometidos por el censado en el momento de responder; o por el censista durante el relevamiento;

c. Después del relevamiento loe errores pueden producirse durante la manipulación física de los cuestionarios, o generarse durante el procesamiento de los datos.

2. Según las causas que los hayan originado los errores pueden clasificarse en:

a. Aquellos que se producen al azar, es decir, sin ningún orden lógico. En caso de que el relevamiento se repitiese en las mismas condiciones, este tipo de errores no volverían a producirse en las mismas variables de los mismos registros. Por ese motivo, se considera que son aleatorios, afectan a todos o a casi todos los valores de todas o casi todas las variables de manera uniforme. Además, se supone que poseen una baja probabilidad de ocurrencia y que no distorsionan, en general, las distribuciones.

Cuando los errores al azar tienen su origen en los individuos censados -en el momento de responder el cuestionario- pueden ser de diverso tipo:

- **Conceptuales**: el informante puede no conocer el concepto acerca del cual se le está preguntando e involuntariamente otorgar una respuesta errónea.
Por ejemplo, el concepto de “trabajo” resulta de difícil tratamiento en el Censo de población. Algunas amas de casa consideran que esa tarea en el hogar debe ser incluida como trabajo. Por otra parte, también existen aquellas mujeres que realizan un trabajo -trabajadores familiares, sin remuneración fija- y que no se declaran como tales en primera instancia, por desconocer el concepto de “trabajo” a que alude el relevamiento. No conocen el concepto que el Censo estipul6 para investigar esta variable.

- **Por errónea interpretación de la pregunta**: el censado puede no comprender en forma correcta qué dato se le solicita e incurrir, en consecuencia, en una respuesta equivocada.
Por ejemplo, puede responder a la pregunta sobre “estado conyugal” contestando en realidad sobre su "estado civil”.

- **Por ignorancia de la respuesta correcta**: se considera que este tipo de error se comete en los casos en los que un individuo informa acerca de las características de una tercera persona.

Por otra parte, los **errores al azar ocasionados por los censistas** también pueden ser de distinto tipo:

- **Conceptuales**: cuando el censista interroga sobre un asunto que no corresponde a lo que en realidad se desea investigar.
Por ejemplo, el caso de censistas que realizaron la pregunta de fecundidad a individuos varones.

- **Por errónea interpretación de la respuesta**: este tipo de error es en especial importante cuando el censista debe completar preguntas con respuestas abiertas;

• **Por erróneo registro de la informaci6n**: Suele suceder que, producto del descuido o el apresuramiento, el censista marque una categoría que no es la que corresponde completar u olvide completar algún dato, o bien podría afectar de manera involuntaria la exactitud de la respuesta transcribiéndola incorrectamente en el cuestionario o no continuando el flujo señalado en el diseño de la cédula censal.

b. Por otra parte, existen otros **errores**, los **sistemáticos**, que se originan en un erróneo tratamiento de la cartografía o de la muestra, o un mal entendimiento de la pregunta, o bien en los conceptos, definiciones o instrucciones preestablecidas, tanto durante el período de organización censal como durante el relevamiento -tanto por parte del respondente cómo del censista- o durante el procesamiento de la informaci6n. Por ejemplo, una mala elaboración de la cartografía puede ocasionar una deficiente cobertura del censo, es decir, que algunas áreas sean omitidas.

Se conoce que en el caso de estos tipos de errores, de repetirse el censo en las mismas condiciones, se producirían muy probablemente en las mismas variables de los mismos registros. Por ese motivo es que suelen afectar a un grupo específico de variables, cuestionarios o registros, y ocasionan cierta distorsión en las distribuciones.

3. **Según el tipo de inconsistencia manifestada:**

Se trata de la forma en que quedan evidenciados los errores en el cuestionario censal.

a. **blanco**: cuando se detecta que falta respuesta en una variable determinada que debía ser investigada en ese individuo;

b. **multimarca**: se verifica más de una categoría marcada como respuesta en la variable;

c. **flujo**: no seguir el flujo señalado en el diseño de la cédula censal, incurriendo en sobremarcas -respuestas en variables que no debían preguntarse a ese individuo- o blancos en variables que debían contener una respuesta válida,

2. Metodología para evaluar y asignar información.

2.1. Características generales.

En general se trata que la etapa de depuración de los datos censales se extienda a todas las etapas involucradas en el censo, es decir durante la fase de preparación, en el relevamiento propiamente dicho y en la etapa posterior al mismo. De esta manera, se considera que se logra detectar errores e inconsistencias de manera eficaz y que se le otorga un tratamiento adecuado.

Por ejemplo, durante la fase preparatoria del Censo 1991 se llevaron a cabo cinco Pruebas Piloto con el fin de ajustar los temas -tipo, número y forma de la pregunta- a investigar en el Censo y diagramar un eficaz diseño del formulario censal .Luego, con posterioridad a la fecha del relevamiento, una evaluación de la cobertura censal permitió evaluar posibles subestimaciones de población.

Por otra parte, una vez captada la información censal, en relación con las variables correspondientes al cuestionario ampliado, el objetivo del trabajo consiste en detectar y posteriormente depurar aquellos errores al azar originados por el entrevistado o por el censista. En este sentido, en este documento sólo se hace referencia a la etapa de consistencia de la informaci6n que se lleva a cabo para detectar errores y corregir los mismos una vez realizada la lectura de los datos como estadio anterior a la obtenci6n de los tabulados con datos del formulario ampliado.

Para realizar la tarea de manera eficiente es necesario definir en forma previa, a partir de los conceptos a que aluden las variables investigadas y las instrucciones impartidas a los censistas, una serie de pautas que posibiliten estimar la calidad de la información censal y llegar a corregir posibles inconsistencias.

La evaluación necesita ser avalada empíricamente. Por ello es necesario analizar cruces especiales con el archivo sucio, es decir el archivo original, para cada una de las temáticas en estudio con el fin de calcular porcentajes de respuestas en blanco/multimarca/error de flujo e inconsistencias[3] (Cuadro 1)

El análisis de los cuadros solicitados contribuye a la definición preliminar de pautas para el tratamiento de la información que presenta algún tipo de inconsistencias. Sin embargo, en la mayoría de los casos requiere de una profundización de las características del error a partir de la elaboración de nuevos cruces de datos. Reuniones periódicas y consultas a diversos especialistas temáticos[4] coadyuva a conformar el marco teórico a partir del cual se estipulan las posibles soluciones.

Para cada temática en particular se elaboran:

1. **Pautas para la detección de errores e inconsistencias** y

2. **Pautas de asignaci6n de información** (determinísticas o probabilísticas) para resolver cada problema en particular.

.Se examinan los casos de no respuesta/multimarca para cada una de las variables, así como la información que, de acuerdo con las pautas establecidas, resulte incoherente.

.Se analiza en especial la dimensión del fenómeno y sus posibles causas.

.Se determina una solución para cada caso específico, en virtud de la premisa de efectuar cambios mínimos y obtener a la vez información coherente.

.Por último, simultáneamente con la redacción de pautas específicas para cada una de las variables tratadas, es necesario construir un diagrama que exprese el texto escrito en términos de programación con el fin de que sea utilizado con posterioridad por el programador.

2.2. Pautas para la detecci6n de errores e inconsistencias.

Para detectar posibles errores en la captaci6n de la información del Censo Nacional de Poblaci6n 1991, se elaboraron en general dos tipos de pautas. Unas, denominadas **de aceptaci6n,** consisten en condiciones lógicas o aritméticas que deben ser satisfechas por los datos para que puedan aceptarse como correctos. Las otras, llamadas **de rechazo o conflicto** corresponden a aquellas condiciones lógicas o aritméticas que, de producirse, provocarían que los datos fueran calificados de err6neos o incoherentes. Las dos se complementan mutuamente en la tarea de detectar posibles errores o inconsistencias en el contenido de la información.

Con el fin de organizar la tarea, se delimitaron 3 (tres) campos específicos en los cuales localizar los errores:

[3] Para el análisis de las variables del cuestionario ampliado se procesó la información correspondiente al distrito XXI de Capital Federal, el Partido de Florencio Varela y, para algunos casos específicos, otros distritos de provincias como la de La Rioja y Formosa.

[4] Fueron consultados especialistas del área de Educación, Minoridad, Salud, Previsión Social, Vivienda y otras áreas del INDEC, así como demógrafos especializados, en determinadas temáticas.

1. En la variable: consiste en determinar si los valores de cada variable captada en forma individual, es decir, sin considerar sus relaciones con valores de otras variables diferentes, pueden aceptarse como correctos. La base para la definición de los códigos de validación se obtiene a partir del diseño de la cédula censal y las instrucciones impartidas a los censistas. En el trabajo de validación se efectúa un control de rangos especificados en forma previa y se comprueba si el valor que asume la variable corresponde a uno de los códigos especificados.

2. En el registro: para cada registro, se comprueba cierto tipo de relaciones -aritméticas y/o lógicas- entre valores variables de un mismo registro. Por ejemplo, hay estados conyugales que resultan ser imposibles o, por lo menos, muy poco probables a determinadas edades para un individuo. En efecto, la situación de un viudo de tres años es a las claras eminentemente errónea.

Existen otras situaciones en las cuales el límite entre la validez y el error de la informaci6n es difuso. Por ejemplo, dato de un individuo de 16 años y estado conyugal viudo posee una probabilidad muy alta de ser erróneo, pero es necesario pensar que si se clasificasen sistemáticamente como erróneos todos los casos de viudos con 16 años, tal vez se perdiesen casos interesantes para el estudio de la población.
Como el tipo de relaciones lógicas o aritméticas entre los valores de las variables de un registro son especificadas por el personal abocado a esta labor, só1o pueden detectarse como anacrónicas aquellas relaciones establecidas por este último.
Ello provoca la necesidad de poseer un riguroso conocimiento acerca de las características de la población en un determinado momento y lugar, con el fin de no introducir mediante procedimientos de asignaci6n más errores que los cometidos en forma previa por el censista y el censado. De allí la necesidad de adoptar medidas concensuadas entre diferentes áreas del organismo e incluso realizar consultas con especialistas externos.

3. Entre registros: los valores de las variables de cada registro conforman una unidad lógica con otros registros de la misma unidad, que en este caso específico se considera el hogar. Por ese motivo, se considera que unos registros y otros han de ser compatibles entre sí. En la depuración de los datos se tiene en cuenta la presencia o ausencia de determinados tipos de categorías de la variable dentro de cada unidad lógica, es decir del hogar. Por ejemplo, determinadas relaciones de parentesco de los miembros del hogar deben guardar coherencia respecto de las relaciones conyugales de los miembros del hogar.

Por ejemplo, a partir de los datos del cuadro 1, en primer término se realiza la **consistencia externa de la informaci6n**, es decir se verifica los resultados obtenidos en el cruce solicitado respecto de la que emanan de otros organismos. Por ejemplo, se confirma la cobertura de salud de la población obtenida a partir del Censo respecto de información de1 Ministerio de Salud.

Luego se especifican pautas en cada uno de los campos con el fin de detectar errores e incoherencias en la información:

1. En la variable: rangos válidos de estado conyuga1:

1. Unido
2. Casado/a en unión legal
3. Separado/a de unión o matrimonio
4. Divorciado/a de matrimonio
5. Viudo/a de unión o matrimonio
6. Soltero/a nunca unido/a
9. Ignorado

2. **En el registro:** la edad debe ser mayor o igual a 14 años.

3. **Entre registros:** Si existe cónyuge del jefe y están los dos presentes en el hogar: ambos deben tener el mismo código en estado conyugal.

Este código debe ser:

1 (unido/a) ó,
2 (casado/a en unión legal) ó,
9 (ignorado).

.A continuación, se tiene en cuenta si existen registros de menores de 14 años con respuesta en la variable.

.Luego se calcula la proporción de respuestas en blanco y con multimarca respecto de la poblaci6n de 14 años y más, es decir aquellos registros que no cumplen con la pauta establecida. En este caso, el total de blancos y multimarca tomados en conjunto en la variable estado conyugal para el jefe del hogar (un universo más acotado del total de poblaci6n de 14 años y más) es sólo de O.7% del total de jefes registrados. Por su parte, la proporción para las cónyuges es de 1.2% respecto del total de censados con dicha relación de parentesco.

.Además, el análisis del Cuadro 1 permite observar incoherencias en la declaración del estado conyugal de los jefes y sus cónyuges censados en el mismo hogar. Se observa que lo que en realidad declaró el censado fue su estado civil y no su estado conyugal. Por ejemplo, el 2.6% del total de jefes censados junto a sus cónyuges declararon un categoría diferente a la de unido o casado y en el caso de los segundos la proporción es de 1.6%. Existen casos en que uno de los dos, o el jefe o el cónyuge declararon unido o casado en el estado conyugal, en tanto no hay una respuesta válida en el otro componente de la pareja.

2.3. Pautas elaboradas para asignar informaci6n.

El objetivo perseguido mediante la elaboración de estas definiciones es el de especificar las posibles soluciones para cada caso concreto, con el fin de no introducir mediante una depuración errónea y encadenada de los datos, mayor cantidad de errores que puedan afectar la calidad de la información.

En general, las decisiones tomadas aluden a asignar información mediante asignación determinística, o bien, probabilística. La primera alude a asignar información de manera automática, por ejemplo en los casos de posibles errores de flujo: a los registros de varones con respuesta en fecundidad se les asigna blanco, pues esa pregunta s6lo debía ser realizada a las mujeres.

La **asignación probabilística** refiere a asignar información a partir de matrices construidas mediante el cruce de dos a cuatro variables independientes con las que se halla asociada, desde un punto de vista teórico, la variable con información inconsistente. Los valores iniciales de la matriz se obtienen a partir de cruces especiales elaborados con información del CEN8O, o de datos válidos y consistidos del Censo 1991 o EPH. La matriz se actualiza constantemente durante el procesamiento de los datos mediante la incorporación de códigos válidos que presentan información coherente mediante la aplicación del método "hot deck"[5].

En aquellas variables en las cuales la proporci6n de blancos es elevada y de acuerdo con el principio de realizar cambios mínimos con el fin de evitar errores en cascada, se opta por asignar código "desconocido" a dicha variable. Por ejemplo, se han considerado este tipo de asignaciones en los casos de las preguntas referidas a "hijos nacidos vivos durante el último año", "jurisdicción" para los obreros o empleados del sector público y "tamaño del establecimiento" para los obreros o empleados del sector privado. Esta última decisión se basa en haber comprobado la baja incidencia de estos casos en relación con el total de variables investigadas en el cuestionario ampliado.

Para todos los casos en estudio se fundamenta cada una de 1as decisiones tomadas.

[5] "Hot deck" es un método de asignación automática de información que consiste en asignar la información del último registro al registro con información inconsistente, a partir de una serie de variables seleccionadas a priori por el investigador y asociadas desde un punto de vista teórico con la variable a consistir.

Una forma de ejemplificar los aspectos teóricos mencionados, se presenta a partir del análisis del Cuadro 1:

.La consistencia de la información se realiza mediante **asignación determinística:**

.para los menores de 14 años con respuesta en estado conyugal:

-se asigna a todos “blanco” en esta variable (1).

Jefe (con cónyuge presente) con multimarca o blanco:

1. Si el cónyuge tiene código 1, 2 o 9 en estado conyugal:

-asignar igual código al Jefe (2).

2. Si el cónyuge tiene código 3 a 6:

-asignar código 1 (uno) al jefe y al cónyuge.

.Además, la **asignación probabilística** refiere a:

Si el cónyuge tiene multimarca o blanco:

-asignar por hot deck código 1 (uno) o 2 (dos) al jefe (matriz para jefes con c6nyuges por sexo y edad). Al cónyuge se le asigna el mismo código que al jefe.

Cada una de las decisiones estipuladas se encuentran fundamentadas. Por ejemplo, en (1) la pregunta sobre estado conyugal sólo debía realizarse a los individuos de 14 o más años. Por otra parte, en (2) se decidió asignar el código “unido” y no asignar por hot deck. El supuesto es que, de existir un código válido 1 (uno) o 2 (dos) en el estado conyugal de alguno de los cónyuges éste debía ser tomado como válido. Al no ser éste ni 1 (uno) ni 2 (dos), o que no coincidan (códigos 1 o 2) entre ambos estados conyugales, se decidió asignar “unido” bajo el supuesto, basado en la experiencia adquirida a partir de las diferentes Pruebas Piloto, que la población tiende a declarar el estado civil y no el conyugal. Empíricamente ello se verifica también a partir del Cuadro 1.

Por último, una vez consistida la información se solicitan tabulados cuyos resultados se comparan respecto de los del archivo sucio, con el fin de verificar la validez y coherencia de los datos consistidos.

Cuadro 1

Estado conyugal – LA RIOJA

		Estado del jefe							
	Total	**Blanco**	*****	**Unido**	**Casado**	**Separado**	**Divor-ciado**	**Viudo**	**Soltero**
Total	16769	105	23	2247	13951	176	47	73	147
Estado del cónyuge	184	10		26	138	2		2	6
Blanco									
*****	19	1	16	1	1				
Unido	2397	19	5	2154	74	73	26	18	28
Casado	13895	74	1	30	13715	22	3	22	28
Separado	101	1		16	14	44	1	4	21
Divorciado	17			3		3	6		5
Viudo	24			2		7	2	10	3
Soltero	132		1	15	9	25	9	17	56

BIBLIOGRAFIA

INDEC. III Curso de informática para estadísticos. Módulo V. Argentina, Centro Regional del IBI para la enseñanza de la informática (CREI) - Centro Interamericano de Enseñanza de Estadística (CIENES), 18 de Noviembre al 15 de Diciembre de 1987. Buenos Aires, 1987.

INDEC. Censo de Población y Vivienda 1980. Cuestionario Ampliado. Definición de pautas para la corrección automática de preguntas con blancos e inconsistentes. Elaborado por Juana Rosa Carrizo. Buenos Aires, 1988. (mimeo)

POPSTAN. A case study for the 1980 Censuses of population and housing. U.S. Department of Commerce. Bureau of the Census-l980, Washington DC, 1985.

UNITED NATIONS. Manual of methods of estimating population Manual II. Methods of appaisal of quality of basic data for opulation estimates. New York, 1955. (Population Studies, 23)