INSTITUTO TECNOLÓGICO DE COSTA RICA
DEPARTAMENTO DE COMPUTACIÓN
PROGRAMA DE MAESTRÍA

GUÍA METODOLÓGICA PARA EL PROCESAMIENTO DE CENSOS Y ENCUESTAS POR COMPUTADOR

Tesis para optar al grado de
Magister Scientae en Computación

(borrador)

Daniel Antich Montero

Cartago, Costa Rica
1997

RESUMEN

El procesamiento de datos de censos y encuestas por computador presenta características propias que lo hacen diferente a otros tipos de sistemas. Estas características determinan muchos aspectos metodológicos particulares que no han sido suficientemente estudiados ni documentados.

Esta guía contribuye en llenar ese vacío existente, presentando una metodología que permitirá tanto al informático como al investigador enfrentar adecuadamente esta problemática.

Aquí se analizan las actividades desde el propio diseño de la investigación, pasando por las etapas de recolección y procesamiento de los datos para terminar con algunas consideraciones sobre la organización final de los datos.

APROBACIÓN DE BORRADOR DE TESIS

GUÍA METODOLÓGICA PARA EL PROCESAMIENTO DE CENSOS Y ENCUESTAS POR COMPUTADOR

Tribunal Examinador

Luis Rosero B., Ph.D.
Profesional Externo

Celso Vargas E. M.L.
Profesor

Freddy Abarca R., Ph.D.
Profesor

José E. Araya, Ph.D.
Profesor Asesor

Carlos González A, Ph.D.
Director de la Maestría

Abril 1997

AGRADECIMIENTOS

Agradezco al Dr. José E. Araya por su constante y valioso apoyo durante todo el proceso de preparación de este documento. También al Lic. Harry Hernández R. quién leyó buena parte del borrador e hizo importantes recomendaciones.

Varias personas de distintas dependencias colaboraron suministrando información, dentro de las que se puede citar a Jorge Mernies de la FAO, Victoria A. Simmons del IPC del U.S. Bureau of Census y Dulce Castillo de Naciones Unidas.

Además, agradezco también a los funcionarios de las biblotecas del Programa Centroamericano de Población de la Universidad de Costa Rica, y de la Dirección General de Estadística y Censos por su colaboración.

INDICE GENERAL

1. INTRODUCCIÓN 1

2. CONSIDERACIONES GENERALES 7

2.1 Antecedentes Históricos 7

2.2 El proceso de los censos y encuestas 13

2.2.1 Descripción del procesamiento de los censos y encuestas . . . 13

2.2.2. Características particulares del procesamiento de censos y encuestas 17

2.2.2.1 Importancia del caso 17

2.2.2.2 Características de la recuperación de datos 18

2.2.2.3 Información estática 19

2.2.2.4 Otras consideraciones 19

2.3 Los Errores en los Datos 20

2.3.1 Errores e inconsistencias 21

2.3.2 Fuentes de error 22

2.3.3 Origen, detección y corrección de errores 24

2.3.4 Estrategia global de manejo de errores 26

2.4 Resumen de la metodología 29

3. DISEÑO DE LA INVESTIGACIÓN 34

3.1 Descripción General 34

3.2 Formulación de la Investigación 37

3.2.1 Consideraciones teóricas 37

3.2.2 Definición de los resultados requeridos 39

3.2.2.1 Cuadros o tabulados 39

3.2.2.2 Definición de variables 48

3.2.2.3 Otros tipos de resultados 49

3.3 Modelaje de Datos 49

3.3.1 Definición de las variables 51

3.3.2 Identificación de las entidades y sus relaciones 53

3.3.2.1 Definición de las entidades 53

3.3.2.3 Definición del cuestionario 55

3.3.3 Especificación de las pautas de consistencia 56

3.3.4 Llaves primarias 66

3.3.5 Definiciones incompletas 67

3.4 Diseño del Cuestionario 68

3.4.1 Consideraciones iniciales 68

3.4.2 Preguntas y variables 70

3.4.3 Clasificación de las preguntas 71

3.4.4 Tipos de preguntas 72

3.4.5 Orden de llenado del cuestionario 76

3.5 Los procedimientos de codificación 78

3.5.1 Tipos de codificación 79

3.5.2 Codificación asistida por computador 82

4. CAPTURA Y PROCESAMIENTO DE DATOS 84

4.1 Consideraciones generales del proceso 84

4.1.1 Las etapas del proceso 84

4.1.2 Procesamiento por lotes 84

4.1.3 Procesamiento descentralizado 87

4.2 Recolección de datos 88

4.2.1 Organización de la Recolección 89

4.2.2 Detección y Corrección de Errores 90

4.3 Crítica/Codificación 93

4.3.1 La crítica 94

4.3.2 La codificación 95

4.4 Entrada de Datos 98

4.4.1 La digitación 98

4.4.2.1 La verificación 101

4.4.2.2 Las características de los programas de entrada de datos 103

4.4.2 Entrada automática 108

4.4.2.1 Comentario General 108

4.4.2.2 OMR: Lectura de marcas ópticas 109

4.4.2.3 OCR: Lectura óptica de caracteres 110

4.4.2.4 Ventajas y desventajas 111

4.5 Detección y Corrección de errores 112

4.5.1 Descripción General del Proceso 112

4.5.2 Tipos de Corrección 113

4.5.3 Corrección Manual 115

4.5.4 Corrección automática 118

4.5.5 Imputaciones 120

4.5.5.1 Asignación Única 120

4.5.5.2 Asignación Variable 121

4.5.5.3 Paquete Caliente ("Hot deck") 123

4.5.6 Otros métodos de corrección 124

4.6 Producción de Resultados Básicos 126

4.6.1 Consideraciones Generales 126

4.6.2 Descripción del proceso 127

4.6.2.1 Tabulación 127

4.6.2.2 Consolidación 131

4.6.2.3 Emisión 132

5. ORGANIZACIÓN DE DATOS 134

5.1 Comentario General 134

5.2 Requerimientos de la organización de datos 135

5.2.1 Captura y depuración 135

5.2.2 Tabulaciones básicas 136

5.2.3 Diseminación 137

5.2.3.1 Datos agregados 139

5.2.3.2 Datos elementales 141

CONCLUSIONES 146

APÉNDICES 149

1. Un Ejemplo de la metodología 150

2. Programas para el procesamiento de censos y encuestas 168

2.1 IMPS: Integrated Microcomputer Processing System 168

2.1.1 Introducción 168

2.1.2 Características básicas 169

2.1.3 Organización de Datos 171

2.1.4 Descripción de los principales módulos 172

2.1.4.1 Entrada de Datos (CENTRY) 172

2.1.4.2 Consistencia y Corrección (CONCOR) 173

2.1.4.3 Tabulaciones (CENTS) 175

2.1.4.4 Recuperación de tablas (TRS) 177

2.1.5 Otras consideraciones técnicas 178

2.2 REDATAM+ 180

2.2.1 Introducción 180

2.2.2 Conceptos Básicos 180

2.2.3 Organización Funcional 184

2.2.4 Estructura interna 186

2.2.5 Otra información técnica 188

3. CASIC: Recolección de información asistida por computador 190

3.1 Introducción 190

3.2 Entrevista asistida por computador (CAI) 191

3.2.1 Entrevista telefónica asistida por computador (CATI) 192

3.2.2 Entrevista personal asistida por computador (CAPI) 192

3.3 Auto-entrevista (CASI) 193

3.3.1 Cuestionario electrónico 193

3.3.2 Respuestas por teléfono de tonos (TDE) 194

3.4 Asistencia del computador en las técnicas tradicionales 194

3.4.1 Lectura de caracteres ópticos (OCR) 194

BIBLIOGRAFIA 196

NOTA BIOGRÁFICA 201

INDICE DE FIGURAS

2.1: Procesamiento de datos de censos y encuestas 14

2.2: Fuentes de error 23

2.3: Revisión y corrección de errores en las etapas del proceso 25

2.4: Componentes del procesamiento de un censo o encuesta 30

3.1: Orden de las actividades en el diseño de la investigación 36

3.2: Diagrama entidad-relación para una Encuesta de Hogares 56

4.1: Características de las etapas del procesamiento 85

4.2: Extremos en la definición de la digitación 99

4.3: Características del programa de entrada de datos 107

4.4: Procesos de detección y corrección de errores 114

4.5: Fases en la producción de resultados 129

5.1: Acceso a únicamente los datos requeridos 144

APÉNDICES

1.1: Diagrama de entidad-relación del ejemplo 159

2.1: Principales módulos del IMPS 170

2.2: Estructura de la base de datos REDATAM+ del Censo de Población y Vivienda de 1982 de Nueva Miranda 183

A cada uno de los miembros de mi familia

que vivió y sufrió de distinta forma

la realización a destiempo de este trabajo.

1. INTRODUCCIÓN

Dentro de la problemática presente en la realización de censos y encuestas, el procesamiento de sus datos por computador es uno de los temas determinantes tanto de la calidad de los resultados como de la eficiencia con que se producen.

Lamentablemente, en muchas ocasiones la metodología aplicada en este tipo de procesamiento no es la adecuada sino que por el contrario, es el producto del desarrollo empírico a lo largo del tiempo, que ha logrado únicamente incorporar algunos avances tecnológicos, o en otros casos, es la utilización de metodologías empleadas en el desarrollo de sistemas de índole administrativo o de control, que presentan características bastante diferentes a las aquí requeridas.

Esta circunstancia ha provocado que el procesamiento de los datos de los censos y encuestas no siempre se realice de manera adecuada y considerando todas las características particulares que deben tomarse en cuenta, y que deben estar presentes en cada una de las etapas de la investigación. Por tal motivo se cometen una cantidad de errores que atentan contra el éxito del proyecto. Dentro de estos errores se pueden citar los siguientes:

Cuestionarios mal diseñados que no permiten determinar bien las entidades y unidades de análisis y que pueden llevar a que las variables recolectas no sirvan

para obtener los resultados esperados.

- Programas de entrada de datos con un enfoque que no contempla las características de captura masiva, datos con errores, verificación, etc.

- Tratamiento incorrecto de los errores e inconsistencias que disminuyen la calidad de los resultados.

- Utilización de software que no se adapta al procesamiento de datos de esta clase de sistemas.

- Organización de los datos finales de forma tal que los procesos de recuperación y consulta resultan bastante ineficientes, o que producen la subtulización de la información.

Esta consideración ha provocado la preocupación entre las personas interesadas en la información que se produce y aún de los mismos directores y profesionales que participan en este tipo de investigaciones. Y no es para menos, pues los recursos que se invierten y el tiempo que se debe esperar para tener los resultados justifican plenamente dicha preocupación.

Lo anterior es probable que tenga como una de sus causas la falta de literatura sobre el tema. Son pocos los documentos que tratan correctamente sobre el procesamiento de censos y encuestas por computador y en la mayoría de los casos enfrentan los principales temas computacionales de manera superficial.

Por tal razón, esta guía se presenta como una contribución para llenar parte del vacío existente presentando una metodología propia, que ayudará al usuario a seguir un orden adecuado en las actividades y a tomar en cuenta todas las consideraciones necesarias que lo lleven a realizar una investigación con una buena posibilidad de éxito.

En otro orden de cosas, esta guía espera servir de referencia tanto al informático que se enfrenta a este tipo de problemas como también brindar información al investigador o profesional en otras áreas que necesita conocer acerca del tema; principalmente a los directores de proyectos de investigación que tendrán en esta guía la información necesaria para comunicarse de manera adecuada con el informático.

A pesar de que este documento se refiere en general al procesamiento de censos y encuestas es conveniente delimitar cuidadosamente el ámbito en donde se espera sea de mayor utilidad.

En primer lugar se debe hacer la diferencia entre las características que presentan los países industrializados con las del resto del mundo. Por una parte, el desarrollo

tecnológico alcanzado ha hecho que en ciertos casos los censos no sean necesarios y en su lugar se utilicen otros sistemas de registro. Por otro lado, la gran disponibilidad de recursos ha modificado los sistemas de captura de datos (que condicionan gran parte del procesamiento) y por ejemplo, en varios países se están haciendo encuestas y otras investigaciones que basan en nuevas técnicas que se apoyan en el uso del teléfono como CATI (Computer Assisted Telephone Interviewing), o con computadores de mano como en CAPI (Computer Assisted Personal Interviewing) o el CASI (Computer Assisted Self Interviewing) en donde el informante obtiene un cuestionario electrónico por disquete o descargándolo de alguna red, lo completa y retorna a la oficina correspondiente.

Por su parte, los países subdesarrollados han mantenido técnicas de procesamiento "tradicionales" pues es poco probable que la mayoría de la población disponga de teléfono para el CATI o todos los encuestadores puedan realizar su trabajo con una computadora de mano (CAPI). Por lo tanto, esta guía tendrá como usuario principal a las entidades de los países subdesarrollados que seguirán usando por mucho tiempo los sistemas de captura tipo PAPI (Paper and Pencil Interviewing). Sin embargo, en el Apéndice 3 se presenta una descripción a manera de referencia de estas nuevas técnicas, conocidas como CASIC[1].

Otro aspecto importante es que a pesar que en el documento se hace mención a los censos y encuestas es claro que el ámbito de la guía va bastante más lejos, y

---

[1]Computer-Assisted Survey Information Collection [US Bureau of the Census-94]

probablemente será de utilidad para una gran cantidad de tipos de investigación más pequeñas que manejen datos estadísticos y similares.

La última consideración que define el ámbito de esta guía se refiere al contenido, pues aquí se tratan únicamente los aspectos computacionales propiamente dichos que van desde el modelaje de datos en la formulación de la investigación hasta el diseño de los sistemas de recuperación y diseminación de información. Esto debido a que dentro del procesamiento de datos de censos y encuestas se puede incluir una otra cantidad de actividades como por ejemplo, planificación de la labor informática, adquisición de tecnología de hardware Y software, capacitación, organización y control del desarrollo de sistemas y del personal de captura de datos, crítica y codificación, etc. que no son tratados en este documento.

Es conveniente aclarar que algunas técnicas propuestas en esta metodología han sido preparadas de tal forma que se adapten fácilmente al IMPS[2] y al REDATAM+[3], dos paquetes muy populares que abarcan todas las áreas del procesamiento descrito en esta guía y que son presentados como ejemplo en el Apéndice 2. De esta forma el usuario podrá utilizar la metodología propuesta aquí, e implementar con cierta facilidad las aplicaciones con estos paquetes.

---

[2] Integrated Microcomputer Processing System, paquete desarrollado por US Bureau of the Census para el procesamiento de censos [US Bureau of the Census-95]

[3] REcuperación de DATos para Areas pequeñas por Microcomputador. Manejador de Bases de Datos de censos y encuestas, desarrollado por CELADE/Naciones Unidas [CELADE-91]

La guía está dividida en cinco capítulos incluyendo el primero que es esta introducción. El segundo hace referencia a los aspectos generales del procesamiento de censos y encuestas desde los antecedentes históricos hasta asuntos más específicos como el manejo de los errores y un resumen de la metodología. En el tercer capítulo se trata el diseño de la investigación partiendo de la formulación teórica para luego presentar los temas sobre el modelaje de datos, los sistemas de codificación y el diseño del cuestionario. El cuarto capítulo se dedica a las etapas de recolección de datos, crítica y codificación y captura y depuración de datos, además de la propia emisión de tabulaciones. Para terminar, el quinto y último capítulo trata sobre la organización de los datos. Posteriormente, en los apéndices se presenta un ejemplo de la metodología, un estudio sobre algunos paquetes y un comentario sobre las nuevas técnicas de recolección de datos.

2. CONSIDERACIONES GENERALES

2.1 Antecedentes Históricos

Los problemas en el procesamiento de los datos de censos y grandes encuestas no son recientes, por el contrario, desde hace muchos años han ocupado la atención de directores de los proyectos y de técnicos que han tratado de mejorar tanto la eficiencia en la obtención de los resultados, como la misma calidad de estos.

Existen referencias en que se menciona que desde hace más de cien años se tenían serios problemas con el tratamiento de los datos de los censos de población de los grandes países. Resalta el hecho de que los últimos resultados del censo de los Estados Unidos de 1880 se obtuvieron muy poco tiempo antes de la realización del siguiente censo, por lo se tenían serias dudas acerca de cuando se tendrían los resultados del censo de 1890.

Esta situación provocó la contratación de un inventor de la época, un señor llamado Herman Hollerith que desarrolló un sistema de tarjetas perforadas que se introducían en máquinas clasificadoras y tabuladoras. Esto permitió que los resultados se tuvieran en un plazo muchísimo menor al estimado. Fue tanto el éxito del invento de Hollerith, que sus máquinas fueron utilizadas para el procesamiento de censos de varios países de Europa e inclusive para los censos de Rusia de esa época.

Este hecho marcó un hito, no sólo en lo que refiere a la obtención de los resultados censales, sino también en el procesamiento de datos en general.

Durante los siguientes cincuenta años se continuó utilizando esta tecnología que se desarrolló principalmente en capacidad y velocidad[4], pero muy poco en funcionalidad.

Posteriormente, y ya en el inicio de la era de la computación, se conoce el uso de una de las primeras computadoras de uso general, la UNIVAC I por parte de la Oficina de Censos de los Estados Unidos para el procesamiento de los censos de 1950. Esta fue la primera agencia civil en utilizar este tipo de equipo de computación electrónica el cual fue capaz de procesar 4000 ítemes por minuto.

La disponibilidad de equipos de computación electrónica marca el inicio de la "era moderna" en el procesamiento de censos y encuestas. Este momento es bien importante ya que se presentaron una serie de problemas, que en conjunto con las soluciones que se adoptaron a lo largo del tiempo, sirvieron de fundamento para el desarrollo del procesamiento de este tipo de datos y también de varios de los temas que se tratan en este documento.

---

[4] De 250 ítemes por minuto en 1890 a 2000 en 1950 [US Bureau of the Census-95]

Las tarjetas perforadas ya no fueron procesadas por los equipos de "registro unitario" con tableros alambrados. En su lugar los datos fueron grabados en cintas magnéticas para ser leídas en computadores electrónicos.

Las primeras computadoras que se empezaron a usar de manera general para procesar censos y encuestas se caracterizaron por disponer de cintas magnéticas como único medio de almacenamiento secundario. Estos computadores eran sumamente lentas y con muy poca memoria principal (8-16 K bytes).

Estos equipos disponían de lectoras y perforadoras de tarjetas que sirvieron de medio de entrada y salida de datos en conjunto con las impresoras de papel en fórmula continua.

Los programas se introducían al computador en tarjetas perforadas conjuntamente con las instrucciones o tarjetas de control que indicaban al computador, y éste al operador, sobre el montaje de cintas y otros "decks" de tarjetas. El acceso a los datos era estrictamente secuencial.

Posteriormente, en los años 60 se utilizaron las computadoras con discos magnéticos de acceso directo, en donde residían los sistemas operativos, lenguajes de programación y algunos pequeños conjuntos de datos. Una computadora de esta época podía tener un poco más de 100 Mbytes en el conjunto de todos los discos magnéticos. Eran más rápidas

y tenían una memoria principal de no más de 256K o 512K bytes. Se disponía de lenguajes de programación de alto nivel como el FORTRAN y el COBOL.

Sin embargo, y debido que el volumen de datos de los censos era mucho mayor que la capacidad de almacenamiento en disco, el acceso de datos continuó de manera secuencial.

Estas consideraciones determinaron las características de los procesos. Los datos leídos de las tarjetas conformaban un archivo maestro en cintas magnéticas que sufría varios procesos.

Primero se le aplicaba un proceso de detección de errores que producía un reporte de inconsistencias para la revisión manual. Las datos para hacer las correcciones eran perforados en tarjetas que pasaban a formar un archivo de movimientos que se ordenaba de la misma manera que el maestro. Posteriormente, se procesaban ambos archivos produciendo un nuevo maestro actualizado que se volvía a revisar para detectar nuevos errores. Si no era posible esta corrección manual, el mismo proceso de detección realizaba las correcciones automáticas en las cuales el computador asignaba un código correcto a la variable con error.

Cuando el archivo maestro estaba libre de errores se continuaba con la emisión de las tabulaciones básicas. Debido a la cantidad de datos a procesar, se trataba de obtener la mayor cantidad de cuadros en una sola "leída" del archivo. Para tal efecto, el archivo se

clasificaba de acuerdo al orden geográfico del país. Luego, durante el proceso de tabulación se acumulaban los casos para los niveles geográficos inferiores para luego consolidar los otros niveles hasta el total general.

Los archivos secuenciales eran organizados en distintos tipos de registro (tipos de tarjeta) físicamente organizados en un modelo jerárquico. Por ejemplo, los registros tipo 1 correspondían a los datos de vivienda, luego seguían los registros tipo 2 de las personas de esa vivienda, etc.[5]

Debido a la dificultad de estos procesos, generalmente no se obtenían más resultados que las tabulaciones básicas, produciendo una gran subutilización de la información censal que era obtenida a un gran costo.

Durante este período se utilizaron con gran frecuencia paquetes de uso general para el procesamiento de censos como la familia de los CENTS y el CONCOR.

Las tarjetas perforadas fueron sustituidas por nuevos sistemas de grabación y lectura de disquetes. Estos equipos tenían la virtud de trabajar con un medio reusable, pero funcionalmente se diferenciaron muy poco de las tarjetas. Tenían algunas pocas ventajas en las facilidades para la revisión de errores.

---

[5] Estas características del procesamiento de censos y encuestas siguen siendo utilizados en la actualidad, principalmente en países poco desarrollados. Además, muchos de esos conceptos están presentes en esta guía.

Más adelante, en los años 70 las máquinas siguieron creciendo en velocidad de proceso, capacidad de almacenamiento en disco y tamaño de la memoria, además que aparecieron las pantallas o terminales para el acceso en línea y las microcomputadoras. Además se desarrollaron fuertemente los Sistemas Administradores de Bases de Datos.

Para los efectos del procesamiento de censos y encuestas, el desarrollo alcanzado en esa época se dio en dos campos principales.

En primer lugar, el uso de terminales (y posteriormente microcomputadores) permitió la captura de datos por medio de un programa sofisticado que aplicaba revisiones de inconsistencias; esto se tradujo en un aumento en la calidad de los datos y mayor eficiencia en los procesos posteriores.

El segundo punto de desarrollo en ese momento fue la aparición de los primeros sistemas administradores de bases de datos de índole estadística[6] que permitieron aumentar considerablemente las posibilidades de diseminación de la información.

Actualmente, entre los esfuerzos que hacen los países industrializados, se distinguen dos grandes áreas.

---

[6] Como el RAPID de Statistics of Canada [Cotton-80]

Por una parte se han desarrollado mucho las técnicas de recolección de datos asistida por computador, dentro de la cual las más conocidas son CAPI (Computer-Assisted Personal Interviewing) en donde se utiliza un computador para realizar la entrevista, y CATI(Computer-Assisted Telephone Interviewing) que se base en una entrevista telefónica. (Ver apéndice 3).

Y como segunda gran área de desarrollo se encuentra la utilización de la información de los sistemas de registro de los países, complementándola con pequeñas investigaciones en vez de censos o grandes encuestas. Por ejemplo, los sistemas de registro civil de personas contienen una gran cantidad de información y se actualizan con los nacimientos y defunciones, por otra parte, los datos de empleo y residencia se pueden obtener de los registros de contribuyentes de impuestos, etc.

2.2 El proceso de los censos y encuestas

2.2.1 Descripción del procesamiento de los censos y encuestas

El proceso de realizar un censo o encuesta, y que se muestra en la figura 2.1, comienza cuando se parte de los objetivos y requerimientos de la investigación para definir las variables que, en forma de preguntas conforman un cuestionario.

Este cuestionario es aplicado por los encuestadores (o enumeradores) a los informantes para obtener las respuestas en cada caso; esto es el llamado trabajo de campo. Nótese que lo anterior puede variar mucho dependiendo del tipo de censo o encuesta que se está realizando.

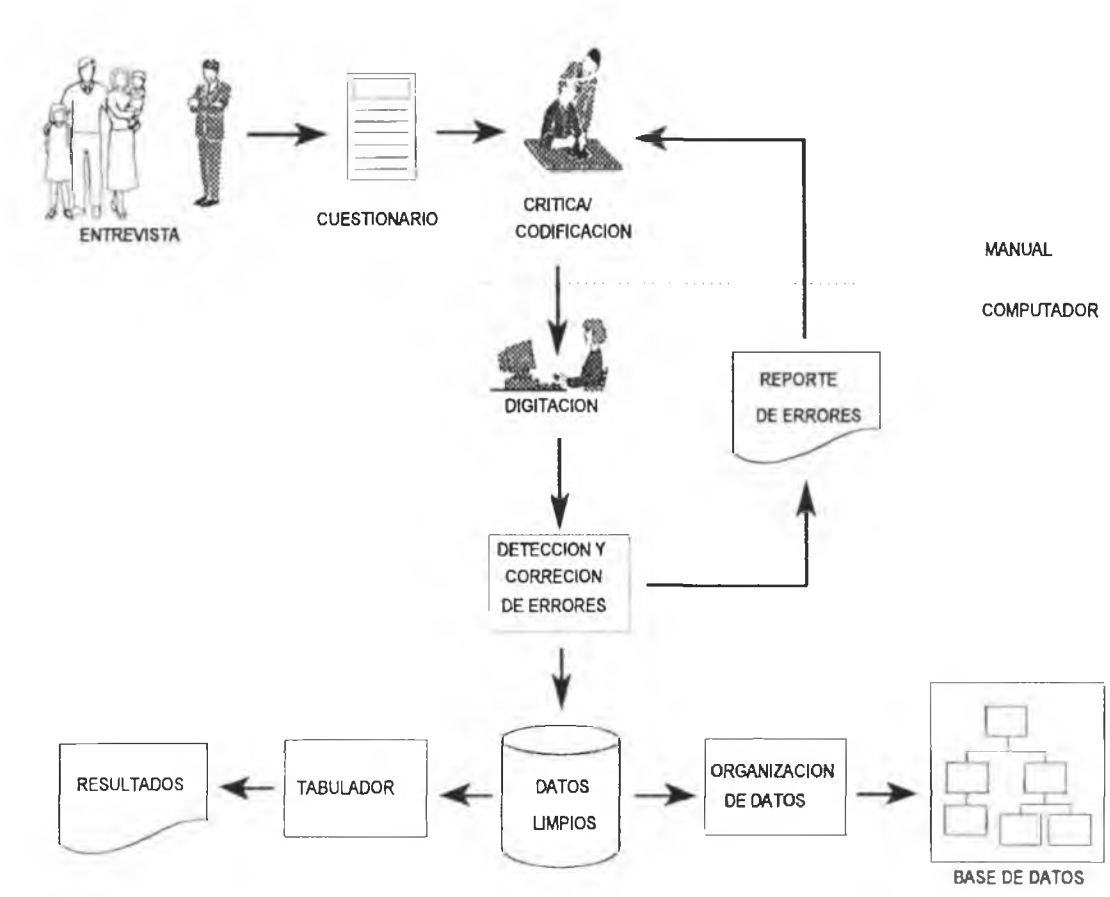


Figura 2.1: Procesamiento de datos de censos y encuestas

Una vez recolectada la información, los cuestionarios con sus respuestas son llevados a nivel central (o regional) para las labores de oficina.

Estas labores de oficina comienzan con la crítica-codificación que se aplica generalmente a la totalidad de los cuestionarios. La crítica consiste en la revisión de algunos asuntos indispensables como la identificación de los datos.

Por otra parte la codificación es la anotación en el cuestionario de algunos códigos para los cuales el encuestador estaba posibilitado a escribir únicamente una descripción; como por ejemplo, la ocupación de la persona.

Después de lo anterior, los cuestionarios están preparados para pasar a la etapa de digitación o en otras palabras, la transcripción de los datos del papel a algún medio legible por el computador. También es común la labor de verificación que consisten en volver a digitar la totalidad o parte de los cuestionarios y variables para tratar de encontrar errores de digitación.

Cuando los datos están en el computador se prosigue con la fase de detección y corrección de errores. Los errores en los datos pueden tener su origen en distintos momentos durante el proceso.

Por ejemplo, el informante se equivocó y suministró información errónea, el encuestador anotó mal la respuesta a la pregunta o el digitador presionó la tecla equivocada. De todas formas, en esta etapa se trata de detectar la mayoría de los errores y efectuar la corrección.

Para tratar los errores se prepara una definición de "pautas de consistencia" que incluye reglas como por ejemplo: si es sexo masculino no puede tener hijos o si es menor de 12 años el estado civil debe ser soltero. Estas pautas son aplicadas a los datos para obtener un conjunto de errores los cuales puede ser corregidos manual o automáticamente. La corrección manual, no siempre posible, implica volver al cuestionario a revisar lo escrito por el encuestador o incluso, hasta regresar al campo a buscar el informante y tratar de encontrar la verdadera respuesta. La corrección automática, implica la imputación de valores de acuerdo a las respuestas de otras preguntas, como por ejemplo si el sexo del jefe de familia es hombre el sexo del cónyuge debe ser mujer. Evidentemente la técnica de imputación automática altera los datos y puede causar un sesgo importante.

Posteriormente, y cuando se dispone de un conjunto de datos "limpios" se continúa con la emisión de las tabulaciones básicas (definidas en el diseño de la investigación). Estas tabulaciones se obtienen para la totalidad del universo y en muchos casos también para agrupaciones geográficas determinadas. Generalmente se requiere recorrer todo el conjunto de casos para calcular los valores de las celdas de cada tabla.

La última etapa consiste en organizar los datos en un sistema tal que permita a cualquier usuario potencial efectuar sus consultas y obtener de manera eficiente los resultados deseados.

2.2.2. Características particulares del procesamiento de censos y encuestas

No sólo su participación en la historia del procesamiento de datos hace de los censos y encuestas un tema particular, sino también sus propias características que difieren significativamente de otros tipos de sistemas. Una de las características más importantes es el manejo de los errores en los datos, que debido a su trascendencia se presenta de manera separada en la sección 2.3 de esta guía. A continuación se comentan algunas otras consideraciones:

2.2.2.1 Importancia del caso

A diferencia de los sistemas de tipo administrativo, en donde cada ocurrencia de una entidad tiene sentido en sí misma, en los censos y encuestas la ocurrencia (o caso) solamente tiene importancia en conjunto con otros casos. Por ejemplo, en un sistema de inventarios, una entrada a bodega es una ocurrencia que tiene relevancia propia pues dice que en tal fecha ingresó tanta cantidad de determinado artículo; por el contrario un

registro censal que diga: masculino, de 35 años, casado, agricultor, y que vive en determinado lugar no da utilidad alguna.

Esta consideración determina principalmente la forma de captar y depurar la información. Por ejemplo, si se encuentran casos con la variable SEXO sin una respuesta correcta, es posible asignarle valores de hombre/mujer de manera alternativa sin afectar los resultados. Incluso, es factible eliminar casos con información equivocada y tampoco tener consecuencias significativas. Por el contrario, y siguiendo nuestra analogía con los inventarios, es prácticamente imposible aceptar una transacción sin un tipo de movimiento correcto y mucho menos eliminar un registro.

Lo anterior ha obligado a desarrollar técnicas para la detección y corrección de casos en las que se hacen imputaciones manteniendo la validez estadística de la información.

2.2.2.2 Características de la recuperación de datos

La gran mayoría de las consultas a los movimientos de un sistema de inventarios requieren todos o casi todos sus atributos, como por ejemplo la fecha, el tipo de movimiento, la cantidad, el artículo, etc. Por el contrario en un sistema de índole estadístico (como los censos y encuestas) una consulta generalmente requiere unas pocas variables. Por ejemplo, si se necesita conocer sobre el estado civil de las mujeres en edad

fértil de cierto lugar, es necesario únicamente tener acceso a la edad, el sexo, estado civil y la residencia y no importa el contenido de las otras variables.

Esta consideración afecta en gran medida el almacenamiento físico y recuperación de la información. Debido a lo anterior, se han desarrollado varios Sistemas Administradores de Bases de Datos que organizan la información de forma "invertida" permitiendo que en un acceso a disco se recuperen muchas ocurrencias de un solo atributo y no todas los atributos de unas pocas ocurrencias, incrementando en gran forma la eficiencia de los procesos.

2.2.2.3 Información estática

También se puede hacer referencia a la característica "histórica" de la información censal y en general de los datos estadísticos ya que es información estática que se refiere a un momento dado en el tiempo y que no es sujeto de actualización. Además, se procesa posteriormente al suceso del hecho y no como un "sistema en línea".

Como consecuencia de esto se tienen muy importantes factores en la forma de almacenar la información y en los medios (por ejemplo CD-ROM) en que se almacena, pensando principalmente en la recuperación y no en la actualización.

2.2.2.4 Otras consideraciones

Unido a lo anterior, se debe mencionar que generalmente las investigaciones censales involucran una gran cantidad de información, por ejemplo en el caso de los censos de población de casi la totalidad de los países se requiere procesar varios millones de registros.

Otra característica importante es la vida útil de los sistemas, pues en la mayoría de las encuestas y censos los sistemas se utilizan una única vez desechándolos posteriormente, asunto que hace tener consideraciones especiales en materia de diseño.

Por el contrario, los datos producto de una investigación de este tipo tienen valor permanentemente pues representan las características de un momento dado de la situación que se está investigando, a pesar de que los sistemas con que se procesan sean desechados.

2.3 Los Errores en los Datos

La necesidad de manejar datos con errores es una de las principales características que hacen del procesamiento de censos y encuestas un campo particular dentro la informática. Las bases de datos de este tipo de sistemas pueden contener ocurrencias con valores fuera del dominio de las variables.

Evidentemente esta condición de error es transitoria y debe ser corregida antes de que se obtengan los primeros resultados.

Un inadecuado manejo de los errores va a determinar negativamente la calidad de los resultados de la investigación, por lo se deben tomar todas las acciones necesarias para obtener un conjunto de datos lo más preciso posible. Estas acciones deben estar presentes en todas las etapas del procesamiento.

2.3.1 Errores e inconsistencias

Es necesario aclarar la diferencia entre el significado de los conceptos de error e inconsistencia. Una inconsistencia es la condición que se da cuando las respuestas de una o más variables tienen un conjunto de valores que no son válidos; por el contrario, un error es un valor de una variable que no es el correcto en el mundo real. Las inconsistencias son las manifestaciones de los errores. Por lo tanto, para corregir las inconsistencias es necesario descubrir y corregir los errores. En muchas ocasiones, el error que produce una inconsistencia no es tan claro, y es necesario analizar el cuestionario para encontrar la correcta causa de la inconsistencia; en otras circunstancias, los errores no producen inconsistencias y es imposible detectarlos.

Por ejemplo, a un caso con sexo femenino se le declaró erróneamente la edad en 12 años, producto de una transposición de dígitos siendo lo correcto 21 años. Es probable

que este error no genere una inconsistencia por lo que no pueda ser detectado, sin embargo, si presenta otra característica como tener hijos, entonces si se produce una inconsistencia ya que las mujeres de 12 años no pueden tener hijos.

2.3.2 Fuentes de error

Los errores en los datos, que provocan las inconsistencias, tienen su origen en distintas fases de la investigación. Estos errores se pueden clasificar por su fuente de la siguiente manera:

errores de diseño: este tipo de error es originado por fallas en el diseño o preparación de la investigación. Aquí se encuentran fuentes como una mala definición de las variables, un incorrecto diseño del cuestionario, instrucciones (manuales) de campo/oficina con fallas, programas de computador que alteran los datos, etc. Estas fuentes de error son sumamente graves y pueden incluso hacer fracasar la investigación. Pero, afortunadamente si hay un cuidadoso control de la preparación de la investigación no se deben presentar estos errores, por lo que esta fuente de error no es tratada en esta guía.

errores de proceso: este es el típico origen que provoca la mayoría de los errores que se presentan en las investigaciones. La figura 2.2 muestra las distintas fuentes de los errores de este tipo en cada una de las etapas del proceso de la investigación.

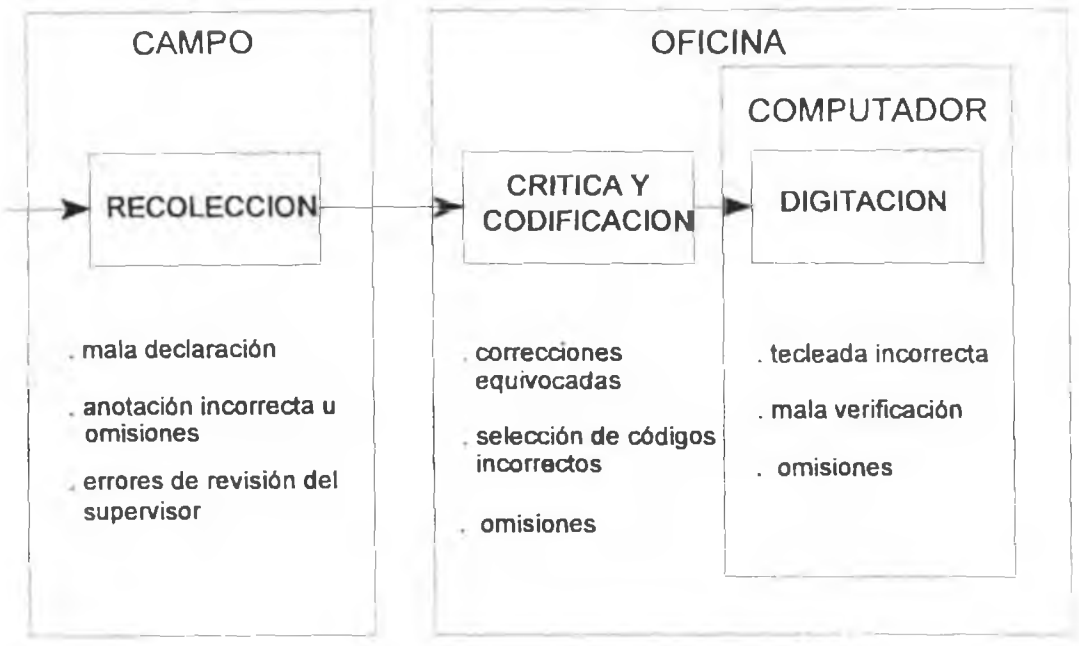


Figura 2.2: Fuentes de error

Durante el trabajo de campo las fuentes de error son: una mala declaración de parte del informante, o una mala anotación de la respuesta de parte del enumerador; también son posibles errores del supervisor de campo (si lo hay) al revisar los cuestionarios. En el trabajo de oficina, los errores provienen de una codificación equivocada o una corrección de datos improcedente. Y la última fuente de error de este tipo se da durante la digitación o entrada de datos al computador cuando el operador introduce un dato que no corresponde a lo reportado en el cuestionario.

2.3.3 Origen, detección y corrección de errores

De lo anterior se concluye que existen tres eventos durante los cuales se pueden generar los errores que son: la recolección que incluye la entrevista y revisión, la crítica/codificación y la digitación.

Para lograr un exitoso tratamiento de los errores, son necesarias algunas actividades en todas las etapas del procesamiento de los datos; desde la preparación de la investigación hasta los últimos pasos de corrección. Es recomendable entonces, definir una *estrategia global de manejo de errores*, que incluya todas las acciones que se deben tomar en cada etapa y la relación entre ellas. Esta estrategia se describe en la siguiente sección.

La figura 2.3 muestra las actividades de detección/corrección de cada etapa, sus beneficios y desventajas.

Nótese que no todas estas actividades necesariamente deben estar presentes en todas las investigaciones. Las actividades de detección/corrección de errores en los datos que se apliquen y sus propias características, deben estar de acuerdo con varios aspectos de la investigación como la cantidad y complejidad de los datos, la disponibilidad de recursos (incluyendo el tiempo), los requerimientos de precisión y calidad, etc.

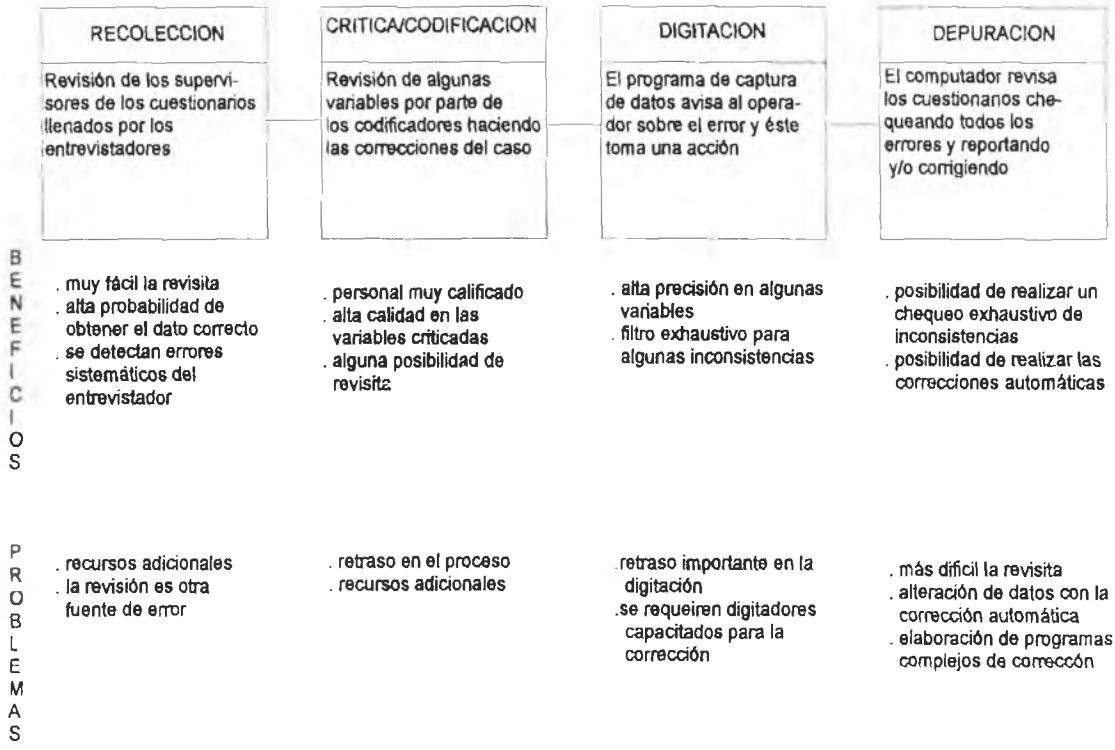


Figura 2.3: Revisión y corrección de errores en las etapas del proceso

De acuerdo al origen del error y a la etapa en que se detecta, así tendrá que ser su corrección. Cuando el error es detectado en la misma etapa en que se produce, su corrección se puede hacer reemplazando el o los valores equivocados con los correctos ya que se está en el evento en que se produjo el error. Por ejemplo, cuando el programa que se usa para la captura de datos reporta un error, el operador puede determinar su origen, y si se trata de un error de digitación lo podrá corregir en ese mismo momento.

Por el contrario, cuando la inconsistencia se originó en una etapa anterior, se tienen dos alternativas de corrección. En una se "regresa" a la etapa en donde se originó el error para determinar el valor correcto y efectuar la corrección. En la segunda posibilidad, la corrección se hace asignando los valores consistentes de acuerdo a algún criterio que puede ser tanto probabilístico como determinístico. Dentro de este tipo de corrección se encuentran las imputaciones automáticas realizadas por el computador en la fase de depuración de datos.

Se debe destacar que entre más cerca del evento en que se produjo el error se detecta, más éxito se tendrá en la corrección. Por ejemplo, sí el mismo entrevistador determina que el niño de dos años no puede ser hijo de la señora de 70 años como lo declara el informante, podrá hacer la corrección durante la misma entrevista evitando que el error llegue hasta el computador y sea necesaria una imputación o una revisita.

2.3.4 Estrategia global de manejo de errores

Como se detalla más adelante, las pautas de consistencia comprenden el conjunto de condiciones que deberán cumplir los datos de un cuestionario para que sea consistente. Entonces, la estrategia a seguir para el manejo de los errores contempla tanto en que etapa del proceso se aplicará cada pauta de consistencia y de que forma se hará la corrección de los errores encontrados.

De acuerdo a las características de la investigación se repartirá la aplicación de las pautas de consistencia en las distintas etapas. En algunas investigaciones se puede dejar la totalidad de los chequeos para que los realice el computador en la etapa de depuración de datos, en otros casos se puede realizar la revisión de muchas variables en las etapas manuales. Sin embargo, debe existir una estricta relación entre los chequeos efectuados en cada etapa.

A pesar de que muchas pautas se pueden revisar en las etapas de campo y oficina, es muy recomendable aplicar la totalidad de las pautas con el computador en la depuración de datos, ya que la máquina no es capaz de pasar por alto inconsistencias ni de generar errores adicionales (siempre y cuando los programas sean correctos). De esta forma se podrá garantizar la obtención de un conjunto de datos consistente.

A continuación se mencionan las consideraciones de manejo de errores en cada una de las etapas del procesamiento:

Diseño de la investigación	Construcción del conjunto de códigos válidos o dominio de las variables, definición de las entidades y del contenido lógico del cuestionario (Modelaje de Datos)
	Definición de las pautas de inconsistencia
Recolección de datos	Instrucciones al Entrevistador/Supervisor sobre el manejo de errores (Manual del Entrevistador/Manual del Supervisor)

Crítica y Codificación	Instrucciones al Crítico/Codificador sobre el manejo de errores (Manual de Crítica/Codificación)
	Criterios de Supervisión y Control de Calidad
Entrada de Datos	Definición del grado de control de inconsistencias en la digitación
	Verificación (doble digitación parcial/total)
Detección y Corrección de errores	Selección del método de corrección
	Instrucciones de corrección manual
	Construcción de las "Reglas de imputación"

En otro orden de cosas, es común que las investigaciones tengan algunas variables sumamente importantes y que se desea controlar su calidad exhaustivamente. Por ejemplo, si se mide el ingreso de la persona, el mismo entrevistador puede aplicar algunas pautas como la relación del ingreso con el nivel de educación del jefe de hogar y con la calidad de la vivienda, etc. Entonces, el entrevistador podrá detectar y corregir algunos datos durante la misma entrevista. Las variables que forman la llave o identificación del cuestionario generalmente reciben una revisión detallada en varias etapas de su procesamiento.

Una práctica muy útil es la revisión con mucho detalle de algunas variables que podrán ser utilizadas como control en la aplicación de otras pautas, ya que en una pauta de consistencia con dos o más variables puede no saberse con certeza cual es la variable con error. De esta forma, si algunas variables son revisadas en varias etapas, se pueden hacer las correcciones suponiendo que éstas variables son las correctas.

También es importante considerar la forma en que se recolecta la información, pues el diseño del cuestionario condiciona de alguna forma los posibles errores. Por ejemplo, una variable que se obtiene de una pregunta precodificada con los valores 1-Sí y 2-No, en donde el entrevistador marque la opción adecuada, se podrá tener un conjunto de respuestas que incluye la respuesta en blanco, una marca al 1 o al 2 o ambas opciones marcadas.

Estas consideraciones del manejo de errores son estudiadas en cada uno de los capítulos de este documento.

2.4 Resumen de la metodología

En los capítulos siguientes se analizan todas las consideraciones que afectan el procesamiento de los datos en cada una de sus etapas, presentando los pasos a seguir y sus características. Sin embargo, y para fines prácticos, en esta sección se incluye un resumen esquemático de todos los elementos que componen esta metodología.

La figura 2.4 muestra los componentes que participan en el procesamiento de un censo o encuesta, divididos en dos fases, la de preparación y la de ejecución.

Las flechas indican el orden en que se deben realizar las actividades o en otras palabras, la relación de requisitos que debe tener cada una. Se han sombreado las actividades que utilizan el computador.

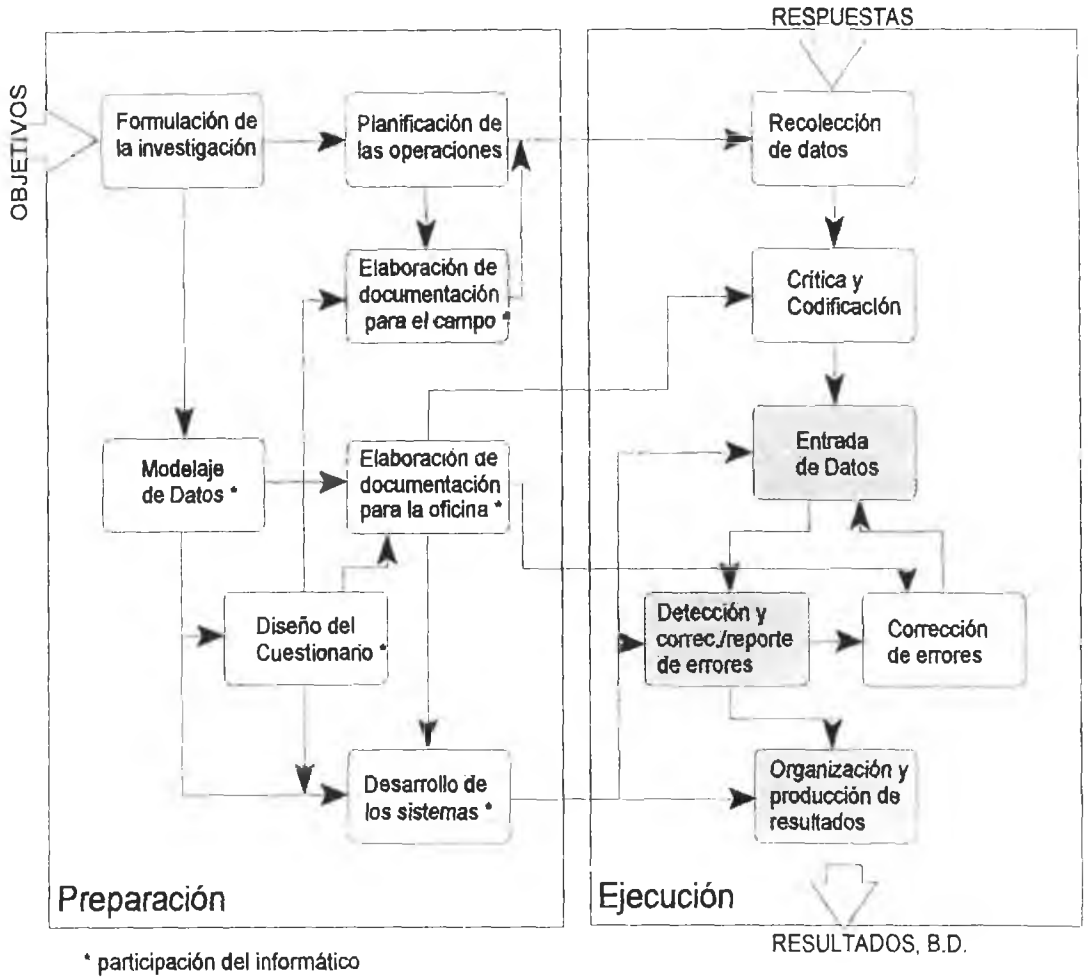


Figura 2.4: Componentes del procesamiento de un censo o encuesta

Para mejorar la presentación de la guía, cada actividad se describe en la etapa del proceso que le corresponde; la elaboración de la documentación de campo se comenta en la etapa de recolección, la documentación de oficina en las etapas de crítica y codificación y en la detección y corrección de errores. Lo mismo sucede con el desarrollo de los programas cuyas características son presentadas en cada etapa.

A continuación se presenta a manera de resumen el contenido de la metodología propuesta.

ACTIVIDAD	PASO	DESCRIPCIÓN
Formulación de la investigación	Definición de resultados	El investigador expresa sus requerimientos en términos de cuadros de salida en un formato especial
		El investigador puede completar su definición de necesidades detallando variables y sus características
Modelaje de Datos	Levantar lista de variables	Se analizan los cuadros de salida diseñados y se listan todas las variables encontradas
	Definición de las características de las variables	Para cada variable se define su nombre, tipo y dominio
	Identificación de las entidades	Se debe identificar a que clase de entidad pertenece cada variable
	Generalización de las subclases	Determinar la existencia de subclases y efectuar la generalización
	Definición de las relaciones entre entidades	Establecer las relaciones entre cada clase de entidad y definir lógicamente el contenido del cuestionario
	Construcción de las pautas de consistencia	El informático en conjunto con el investigador, preparan las pautas de consistencia necesarias para controlar la calidad de la investigación
Diseño del cuestionario y sistemas de codificación	Definición de las preguntas	El investigador con apoyo del informático define las preguntas a incluir en el cuestionario
	Creación de las secciones o grupos de preguntas y saltos	De acuerdo a la entidad a que pertenece cada variable y a las subclases existentes se establecen las secciones del cuestionario y los saltos internos

	Tipificación de las preguntas	Se debe determinar cada pregunta a que tipo pertenece
	Definición del sistema de codificación de cada pregunta	Para cada pregunta se debe definir como y cuando se hará la codificación
Recolección de datos	Diseño del sistema de manejo de lotes	Especificar el control primario definiendo las áreas de enumeración, los lotes, el control de calidad y otros consideraciones
	Elaboración de los manuales de enumeración y supervisión	Es necesario incluir las instrucciones de revisión de posibles inconsistencias, tanto en la propia entrevista como en la supervisión
Crítica y Codificación	Elaboración de los manuales de crítica y codificación	Se debe definir el conjunto de revisiones que harán las personas encargadas de la crítica y los procedimientos de corrección
		Para cada pregunta que requiera codificación manual se debe incluir las instrucciones al codificador, además se debe construir o adquirir los manuales de códigos requeridos
Entrada de Datos	Diseño del programa de entrada de datos	Determinar el nivel de revisión de inconsistencias de la entrada de datos, estableciendo cada una de las pautas que serán chequeadas y los procedimientos de corrección
		Establecer los saltos en la secuencia de captura de acuerdo a las secciones del cuestionario
		Definir el tipo de verificación
Detección y corrección de errores	Diseño del sistema de detección y corrección de errores	Definir el tipo de corrección para cada variable
	Elaboración de las reglas de imputación	Incluir cada una de las reglas para las variables sujetas a corrección automática

	Elaboración del instructivo de corrección manual	Detallar el procedimiento de corrección manual asociado a cada pauta de consistencia
Producción de resultados básicos	Determinación del orden del proceso de los datos	De acuerdo con los cortes de área definidos en el diseño de la investigación, determinar los ordenes en que se deberá procesar el archivo de datos limpios

3. DISEÑO DE LA INVESTIGACIÓN

3.1 Descripción General

En muchas ocasiones, y probablemente como herencia de los procesos manuales se comete el error de realizar las actividades del diseño de la investigación en un orden inadecuado, o de manera muy superficial. Esto trae como consecuencia que los resultados obtenidos sean limitados en calidad y que posiblemente no lleguen a cumplir los objetivos propuestos.

Por ejemplo, es muy común la tendencia de intentar diseñar el cuestionario o instrumento de recolección y luego esperar que el informático obtenga de éste la información necesaria para desarrollar el sistema de procesamiento de los datos [Sierra Bravo-88][Naghi-87]. Más aún, en algunas ocasiones, el contacto con el informático se produce despues del propio levantamiento de la información.

Otro problema común es la definición muy superficial de los requerimientos durante el diseño y que luego, cuando se necesita detallar el contenido de los cuadros de salida, se encuentra el inconveniente de que la información que se obtiene del cuestionario no permite producir el cuadro exactamente como se desea.

En este capítulo se presenta una propuesta con la descripción de las actividades necesarias para el buen diseño de una investigación, el orden adecuado y sus características. En la figura 3.1 se muestran estas actividades de diseño en el orden propuesto en esta guía.

Como actividad inicial aparece la propia formulación de la investigación, en donde se debe realizar todo lo relacionado con la definición de los objetivos de la investigación, el establecimiento de las hipótesis, el marco teórico correspondiente, etc. El principal producto de esta formulación consiste en la correcta definición de los resultados que deberá producir la investigación.

En la definición de los resultados, el investigador estará obligado a preparar adecuadamente sus solicitudes de tabulados y otros resultados. Con base en estos resultados, el informático podrá hacer su modelaje de datos, definiendo entidades y relaciones, sus variables y los dominios de estas, así también como las pautas de consistencia correspondientes.

Una vez completo el modelaje de datos se tendrán todos los requisitos para pasar al diseño del cuestionario y a la definición de los sistemas de codificación. Esta labor efectuada en conjunto con el investigador producirá una boleta como instrumento básico de recolección de los datos.

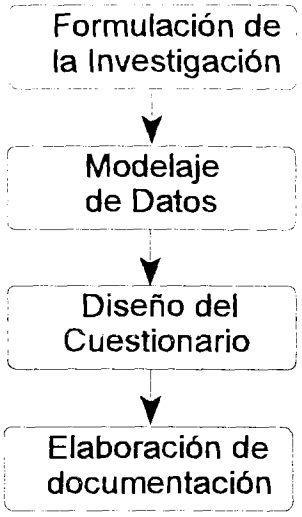


Figura 3.1: Orden de las actividades en el diseño de la investigación

Como última actividad se encuentran la elaboración de los manuales para el trabajo de campo (enumeración/supervisión) y para la labor de oficina (crítica/codificación/corrección), así también como la preparación de las reglas de imputación.

A continuación se detalla cada una de estas actividades.

3.2 Formulación de la Investigación

3.2.1 Consideraciones teóricas

La mayoría de las investigaciones parten de una hipótesis que se plantea como solución a un problema determinado, y por lo tanto su objetivo principal es comprobar o desechar tal hipótesis.

Estos procesos de investigación requieren, en la mayoría de los casos, de datos que sirvan para medir la certeza o no de la hipótesis. Los datos representan el estado de los hechos a investigar en el momento de la observación.

Generalmente, el investigador no dispone de todos los datos que requiere, por lo que deben plantearse estrategias para su obtención; y es en este sentido que se utilizan comúnmente las encuestas y censos como medio de recolección de los datos requeridos.

Cuando una investigación se basa en una encuesta o censo para la obtención de los datos, es indispensable que se consideren los aspectos relacionados al tratamiento de estos datos por computador, dentro del diseño y la formulación de la investigación. De esta forma será posible planificar adecuadamente todas las actividades informáticas y evitar problemas que puedan atentar contra la calidad y hasta con el éxito final de la investigación.

A pesar de lo anterior, esta actividad es de responsabilidad total de los directores del proyecto y profesionales en el campo propio de cada investigación. El informático podrá participar a solicitud de estos para emitir su parecer sobre un tema específico, o para estudiar la factibilidad de determinado proceso.

Desde el punto de vista computacional, el producto principal de esta actividad es la definición de los resultados que se desea obtener de la investigación. Esta definición servirá de base para el diseño de todos los componentes en el subsiguiente procesamiento de la información.

Sin embargo, es necesario considerar que pueden existir un conjunto de resultados que no son requeridos para los fines propios de la investigación, pero que pueden ser de gran utilidad para otro tipo de usuario (probablemente externos a la investigación).

La solicitud de resultados originales de la investigación es satisfecha como parte del proceso principal, y las nuevas consultas y requerimientos deberán interactuar con la base de datos producto final del procesamiento de los datos. Parece interesante que los investigadores traten de predecir las posibles consultas a esta base de datos y preparen una especificación de esos requerimientos. De esta forma el informático tendrá mayor criterio para organizar los datos finales, y el nuevo usuario tendrá también una mejor respuesta a sus solicitudes.

3.2.2 Definición de los resultados requeridos

El investigador deberá diseñar los reportes que se obtendrán al final del procesamiento de los datos. Investigadores con poca experiencia en esta práctica, podrán solicitar la participación del informático para que ambos en conjunto preparen el diseño de los resultados. Estos resultados deberán ser expresados de manera concreta y detallada para que no se presenten definiciones incompletas ni interpretaciones ambiguas.

Es de considerar el hecho de que el informático no siempre conoce sobre el tema de la investigación, razón por la cual, el investigador deberá ser sumamente explícito y detallado en la definición de los resultados.

Nótese que esta construcción de resultados por parte del investigador, no es otra cosa más que la definición implícita de las variables, sus características y las unidades de análisis o "niveles de observación" [Muñoz-89].

La definición de resultados se hace de varias maneras que se detallan a continuación:

3.2.2.1 Cuadros o tabulados

Este tipo de resultado es el más común, principalmente en investigaciones de índole poblacional. Aquí se presentan los datos de forma tabular por medio de filas y columnas.

El diseño de estos cuadros debe ser completo y debe abarcar tanto lo referente al contenido del cuadro como también el detalle de los encabezados nombres de columnas/filas, pie de página, etc. Sin embargo para los fines de esta guía se dará importancia únicamente al contenido del cuadro, con la excepción del título que puede ayudar a completar la definición. El contenido del cuadro se divide en las varios componentes como se muestra a continuación:

título

universo

cortes

columnas

filas		celda	

Título

Para especificar el título del cuadro se acostumbra a usar la siguiente convención:

	Contenido del cuadro
por:	Detalle de las columnas
según:	Cortes y detalle de las filas

por ejemplo:

TOTAL DE VIVIENDAS INDIVIDUALES OCUPADAS
POR: MATERIAL DE LOS TECHOS
SEGÚN: REGIÓN Y MATERIAL DE LAS PAREDES

Esta forma de dar título a los cuadros ayuda a definir sus características. Del "Contenido del cuadro" se puede obtener el universo de casos, del "por" la definición de las filas y del "según" los cortes y las columnas.

Universo

La definición del universo de un cuadro sirve para delimitar el conjunto de casos que se deben tomar en cuenta para el cálculo de su contenido. Esto se hace necesario ya que en una gran cantidad de cuadros no se utiliza la totalidad de los casos de la investigación. Es común que el universo se defina implícitamente en la propia redacción del título del cuadro.

Esta definición consiste en un predicado bastante libre pero que debe ser consistente, como por ejemplo:

mujeres de 15 a 49 años

Cortes

Los cortes, llamados también cortes de área, determinan los subconjuntos de casos para cada uno de los cuales se desea obtener un cuadro similar. Si no se define corte alguno, el cuadro se emitirá una única vez para la totalidad de los casos del universo.

Estos cortes son comúnmente usados para obtener cuadros para cada una de las áreas geográficas o política-administrativas en que se divide un país o región.

La definición de estos cortes que propone esta guía se hace completando una pequeña tabla como la siguiente:

NIVEL	CORTE	SUMARIO
0	TOTAL	S/N
1	Corte 1	S/N
2	Corte 2	S/N
:	:	:
n	Corte n	--

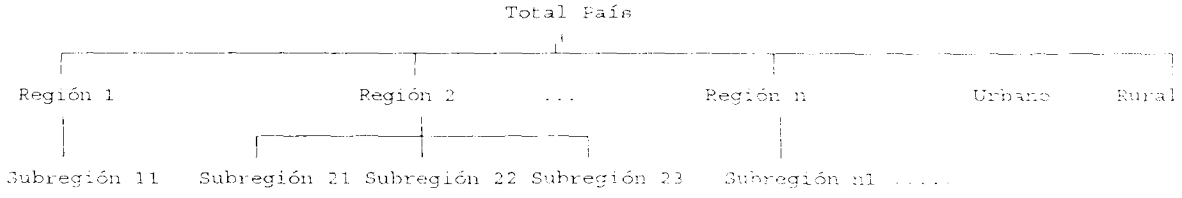
El nivel indica el grado de desagregación en que se presentará el cuadro. Nivel 0 para todo el universo, nivel 1 para la primera desagregación, etc. La columna "CORTE" se puede completar con una palabra o frase que define el concepto de corte. Para el nivel

0, el corte es siempre "TOTAL". Pueden haber varios cortes para un mismo nivel. Por último, la columna "SUMARIO" permite al usuario indicar si se emite o no (S o N) el cuadro resumen para ese corte. Los cuadros SUMARIO se pueden obtener para cualquier nivel n-1. Cuando solamente se define un nivel 0 de corte, es necesarios indicar S para la columna sumario para obtener el único cuadro a nivel total. El orden en que se anoten los cortes en esta tabla tiene relevancia solamente cuando existen dos o más cortes para el mismo nivel, en este caso los cortes de niveles inferiores deben estar inmediatamente después de su correspondiente nivel superior.

Por ejemplo, en una investigación a nivel nacional, es posible solicitar un cuadro con la siguiente definición de cortes:

> para todo el país, para cada una de las regiones, y dentro de éstas, para cada subregión, además, el total del país dividido por zona urbana/rural.

lo que produciría lo siguiente:



La tabla que permite al investigador definir sus cortes quedaría de la siguiente forma:

NIVEL	CORTE	SUMARIO
0	TOTAL	S
1	Zona	S
1	Región	S
2	Subregión	--

Es posible que esta forma de especificar los cortes no sea suficiente para determinadas ocasiones complejas. En esas situaciones, el investigador deberá completar su definición del cuadro con todas las anotaciones necesarias para lograr una especificación satisfactoria.

Filas/columnas

Cada entrada al cuadro recibe el nombre de fila o columna si se trata de una entrada horizontal o vertical respectivamente. Estas pueden ser independientes o dependientes de otras.

Por ejemplo, una columna dependiente puede obtenerse como la suma de otras dos columnas que pueden ser independientes o a su vez dependientes de otras columnas.

La definición de las columnas o filas dependientes se puede lograr fácilmente numerando las filas de arriba hacia abajo y las columnas de izquierda a derecha. Entonces, cada entrada dependiente puede ser especificada anotando la operación y los correspondientes números de entradas. Por ejemplo, en la columna (4) están los hombres, en la (5) las mujeres, entonces, la columna (6) del total se puede definir como (6) = (4) + (5).

Las entradas independientes son más complejas y se requiere que el usuario defina dos características para cada una.

En primer lugar se debe detallar el subconjunto de casos que corresponden a la entrada en particular. Por ejemplo, algunas columnas de una tabla podrían ser: hombres casados, hombres solteros, hombres viudos, etc. Aquí también es probable que el propio encabezado de la fila o columna determine el subconjunto de casos.

Debe existir una estricta consistencia entre el universo del cuadro y el subconjunto de casos de las entrada a la tabla, de tal forma que todo subconjunto pertenezca al universo, y que todos los casos del universo del cuadro tengan su correspondiente fila/columna.

La segunda característica de la definición de una entrada es su propio contenido que puede ser de dos tipos: conteo, acumulación. El conteo indica que cada celda contiene el número de casos que corresponden al subconjunto de su fila y de su columna. La

acumulación es similar al conteo, solamente que la celda tiene la suma de determinada(s) variable(s) de cada caso que corresponda a esa fila/columna.

El tipo "acumulación" es el indicado para las encuestas por muestreo, en las cuales es necesario expandir los resultados de acuerdo a factores de ponderación, que son las variables de acumulación.

Es necesario que exista correspondencia entre la definición de una fila y la columna. Además, si el contenido de todas las entradas es igual, entonces bastará que se indique para todo el cuadro. Por ejemplo, si las columnas de una tabla se refieren a la cantidad de ocupantes de la vivienda, 1, 2, 3, etc, es necesario únicamente detallarlo una vez.

Otra consideración importante son las variables derivadas que no son fácilmente comprensibles. Por ejemplo, si se solicita un cuadro con el nivel de hacinamiento, el investigador deberá especificar detalladamente cual es la fórmula para el cálculo de esa variable.

Celdas

Las celdas son la intersección de una fila con una columna y su contenido está determinado por éstas. Solamente es necesario especificar una celda en particular cuando tenga características distintas a las de su fila y columna correspondientes. Por ejemplo,

es posible que en algún caso se requiera que una celda no sea el total de otras, sino la mediana.

Un ejemplo de cuadro típico en los Censos de Vivienda, se muestra a continuación:

TOTAL DE VIVIENDAS INDIVIDUALES Y COLECTIVAS OCUPADAS
POR: ZONA, TOTAL DE OCUPANTES Y PROMEDIO DE OCUPANTES POR VIVIENDA
SEGÚN: REGIÓN Y TIPO DE VIVIENDA

REGIÓN Y TIPO DE VIVIENDA	TOTAL			URBANO			RURAL		
	VIVIENDAS	OCUPANTES	PROMEDIO	VIVIENDAS	OCUPANTES	PROMEDIO	VIVIENDAS	OCUPANTES	PROMEDIO
	(1) (4)+(7)	(2) (5)+(8)	(3) (2)/(1)	(4)	(5)	(6) (5)/(4)	(7)	(8)	(9) (8)/(7)
REGIÓN 1									
TOTAL	(1)=(2)+(7)								
INDIVIDUAL	(2)=(3)+..+(6)								
CORRIENTE	(3)								
MÓVIL	(4)								
EVENTUAL	(5)								
TUGURIO	(6)								
COLECTIVA	(7)								

Nota: Calcular primero los totales y luego los promedios

Este cuadro presenta 9 columnas, de las cuales únicamente 4 son independientes (4, 5, 7 y 8) y el resto son dependientes. En el caso de las filas, se tienen 5 filas independientes (de la 3 a la 7). El contenido de las columnas y filas independientes queda suficientemente definido en los propios encabezados y literales del diseño de la tabla.

En este caso específico, el orden en que se calculas las entradas dependientes altera los resultados por lo que se menciona explícitamente el orden correcto.

El universo se encuentra en el título del cuadro: VIVIENDAS INDIVIDUALES Y COLECTIVAS OCUPADAS, y los cortes se deben definir adicionalmente de la siguiente manera:

NIVEL	CORTE	SUMARIO
0	TOTAL	S
1	Región	--

La "S" en el sumario para el nivel 0 indica que además de un cuadro para cada región, es necesario obtener un cuadro acumulado para el total del país.

3.2.2.2 Definición de variables

El investigador puede también incluir en los resultados requeridos la definición explícita de variables. Para esto debe aportar la siguiente información:

- nombre y descripción de la variable
- tipo de variable:

  numérica continua o categorizada, alfanumérica o fecha.
- conjunto de valores válidos (dominio)

- categorías y sus nombres
- universo de casos en donde aplica

3.2.2.3 Otros tipos de resultados

Esta metodología para definir un cuadro o una variable no es exhaustiva, debido a que pueden presentarse una gran cantidad de posibilidades que no es factible sistematizarlas. Sin embargo, la gran mayoría de los cuadros o tabulados de un censo o encuesta se pueden definir de esta manera. El resto deberán especificarse utilizando el mismo nivel de detalle y sobre todo el sentido común.

3.3 Modelaje de Datos

Una vez que el investigador ha presentado detalladamente la definición de los resultados requeridos se puede comenzar con el modelaje de datos. Esta etapa forma parte de la preparación de la investigación y debe producir un diseño completo de los datos que se van a procesar posteriormente.

El modelaje de datos puede realizarse por medio del desarrollo de un modelo de entidad-asociación por lo que debe presentar una definición exhaustiva de lo siguiente:

- variable o atributos
- dominio o conjunto de valores de cada variable
- entidades presentes en la investigación
- clases y subclases de entidades
- relaciones entre las entidades
- pautas de consistencia

Nótese que la única fuente de información que tiene el informático para completar el modelaje de datos es la definición de los resultados. Si en algún momento se obtiene información adicional del investigador, debe ser a través de una mejor y más completa especificación de esos resultados requeridos. Esto indica que es posible que la especificación de resultados regrese al investigador para afinar su definición.

El procedimiento para realizar el modelaje puede estar determinado por la habilidad y experiencia del informático en este tipo de actividades, sin embargo, a continuación se proponen una serie de pasos que ayudan a que el resultado del modelaje cumpla con los requisitos anteriores.

Estos pasos son: la definición de las variables, la identificación de las entidades y sus relaciones y la especificación de las restricciones o pautas de consistencia. En el apéndice 1 se presenta un ejemplo completo de esta metodología.

3.3.1 Definición de las variables

En este paso se debe asegurar que se incluyan dentro del modelo de *datos todas las* variables con una definición apropiada que permita obtener posteriormente los resultados solicitados por el investigador. Errores u omisiones en la definición de las variables van ha condicionar negativamente la calidad y eficiencia en la producción de resultados.

Como se ha mencionado, el investigador debe presentar su solicitud de resultados generalmente en forma de cuadros o tabulados, por lo que se parte de estos para el inicio de la definición de las variables y posteriormente se incluye la información de otros tipos de resultados.

En cada una de las partes de un cuadro o tabulado, el investigador ha detallado sus requerimientos por medio de la mención de variables y/o valores de variables. En la mayoría de los casos se presenta explícitamente el nombre de la variable, en algunos otros casos únicamente la descripción de los valores (hombre/mujer en vez de sexo).

Entonces, para la identificación de las variables, el informático debe revisar el diseño de todos los cuadros o tabulados en cada una de sus partes (cortes, universo, filas/columnas, celdas, etc.), levantando una lista de todas las variables y valores presentes. Para los casos en los que se menciona el valor de la variable, será necesario agrupar estos valores de una misma variable para poder definirla correctamente.

Cada variable se define de la siguiente manera:

nombre: debe ser único para identificar plenamente y describir a la variable.

tipo: las variables pueden ser de distinto tipo como numéricas contínua, numérica categorizada, caracter[7] y fecha.

dominio: aquí se deben indicar el o los rangos aceptados, y/o cada una de las categorías, su código correspondiente y un texto que describa su significado.

Para este paso se deben tener en cuenta las siguientes consideraciones:

- cuando no se encuentre explícitamente el nombre de la variable, éste debe ser deducido a partir de los valores presentes.

- es posible encontrar distintos conjuntos de valores para una misma variable, como por ejemplo distintos grupos de edad. En este caso es necesario establecer un conjunto de valores elementales que permita una derivación hacia todos los demás conjuntos.

[7] Es poco común encontrar variables con contenido caracter y no se recomienda su uso ya que no se prestan a un tratamiento de índole estadístico.

3.3.2 Identificación de las entidades y sus relaciones

3.3.2.1 Definición de las entidades

Cada entidad no es otra cosa más que las unidades de análisis o de observación definidas implícitamente por el investigador. Por ejemplo, las entidades en investigaciones poblacionales, pueden ser la vivienda, el hogar y la persona.

En este paso se debe identificar cuales entidades están presentes en la investigación y cual es las relación entre ellas.

La mejor manera de "descubrir" cuales son las entidades de la investigación es estudiar cada una de las variables listadas en el punto anterior y preguntarse para cada una de ellas de que entidad es atributo. Por ejemplo, una de estas preguntas podría ser: ¿De que es atributo el material del piso?, por lo que se descubre que el material del piso pertenece a la vivienda que es una entidad que participa de la investigación. De esto se obtendrá una lista de posibles entidades que debe ser analizadas para determinar cuales son las verdaderos tipos de entidad a incluir en la investigación.

"Las entidades son "cosas" del mundo real que tienen una existencia independiente" [Elmasri-89] por lo tanto se debe tener cuidado en no confundir su definición, y manejar

correctamente el tema de las clases y subclases de entidades como se explica a continuación.

3.3.2.2 Generalización de clases de entidades

En este tipo de sistemas es recomendable la definición de ciertas subclases de entidades y posteriormente hacer la generalización correspondiente para obtener la clase de entidad requerida. Por ejemplo, en investigaciones sobre la población es común que algunas variables tengan sentido únicamente para personas mayores de determinada edad, como las preguntas sobre el nivel de instrucción y la asistencia escolar se hacen a las personas mayores de 5 años, las preguntas sobre la condición laboral y estado civil tienen sentido sólo para las personas de 12 años y más. Entonces, el modelaje de datos debe considerar en el primer momento una clase de entidad llamada "Personas de 12 años o más" con atributos "estado civil" y "condición de actividad laboral", luego se hará una generalización a la clase llamada "Personas de 5 años o más" con atributos como "asistencia escolar", etc y para terminar en la clase de entidad "persona".

Esta definición de clases y subclases tendrá una gran importancia a la hora de diseñar el cuestionario, ya que las subclases pasarán a formar secciones de cuestionario. Adicionalmente esta condición afectará el programa de entrada de datos, ya que las secciones producen los "saltos" en la secuencia de captura.

3.3.2.3 Definición del cuestionario

La definición del contenido lógico del cuestionario corresponde a las relaciones entre los tipos de entidad presentes en la investigación. Esta definición tiene dos características muy importantes:

- organizadas jerárquicamente

- cada ocurrencia de la entidad raíz es un cuestionario

Por ejemplo, en las encuestas de hogares se puede tener una estructura como la mostrada en la figura 3.2.

Entonces, cada ocurrencia de la entidad vivienda define un cuestionario. Además las relaciones entre vivienda y hogar y entre hogar y persona son 1 a N. Cada vivienda tendrá uno o más hogares y cada hogar una o más personas. Evidentemente, las entidades hogar y personas son débiles pues no tienen existencia propia sin una vivienda. Por otra parte, la participación de hogar es parcial, ya que las viviendas desocupadas no contienen hogares.

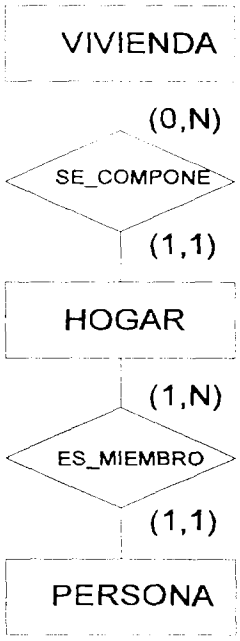


Figura 3.2: Diagrama entidad-relación para una Encuesta de Hogares

3.3.3 Especificación de las pautas de consistencia

Como se ha mencionado anteriormente, una de las características principales del procesamiento de datos de censos y encuestas, es la necesidad de aceptar y manejar datos con errores, principalmente durante las primeras fases del proceso. Esta consideración hace que la definición de las restricciones de integridad de los datos presenten características particulares, y tengan que tratarse con mucho más cuidado.

La teoría sobre base de datos relacional [Elmasri-89] describe las restricciones de integridad como parte del esquema de la base de datos y que deben estar presentes en cada instancia u ocurrencia de ese esquema. Estas restricciones de integridad son:

- La restricción de llave que indica que todas las llaves candidatas de cada instancia del esquema deben ser únicas.

- La restricción de integridad de entidad en que ninguna llave primaria puede incluir valores nulos.

- La restricción de integridad referencial en que cuando una ocurrencia de una entidad hace referencia a otra relación entonces debe referirse a una ocurrencia que exista en esta otra relación.

A esto hay que agregar las pautas de consistencia que determinan si el cuestionario es válido o no. Sin embargo, estas pautas no son restrictivas ya que se deben aceptar todos los cuestionarios aún los que incumplan una o más de estas pautas. Estos últimos tendrán el estado de *inconsistentes*.

Las ocurrencias de cada cuestionario deben ser evaluadas *contra* estas pautas de consistencia. Si la evaluación de todas las pautas resultan correctas, entonces el

cuestionario es válido, por el contrario, si al menos una falla, el cuestionario es inconsistente y debe ser corregido.

Nótese que "Se dice "consistente" y no "correcto" 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" [Ortúzar 78].

A diferencia de las restricciones de integridad que se evalúan y disparan cuando se está actualizando la base de datos, este otro tipo de restricción puede ser evaluada posteriormente en un proceso específicamente diseñado para ese fin. Por lo tanto la base de datos debe tener la posibilidad de aceptar y contener datos erróneos o sea, que no cumplen con las pautas de consistencia.

Existen excepciones que deben tomarse en cuenta, como la posibilidad de cierto tipo de pautas no obligatorias o pautas *suaves*. Este tipo de pauta se puede usar para controlar la gran mayoría de los casos pero existe la posibilidad que algunos casos no cumplan la pauta. Por ejemplo, las mujeres no pueden tener hijos antes de los quince años, pero existe la posibilidad que algunas niñas si los tengan. Por otra parte, pautas obligatorias o *duras* pueden "fallar" con casos excepcionales. Por ejemplo, una vivienda no puede tener paredes de cartón y techo de concreto, sin embargo, alguien que duerma bajo un puente con algunas cajas de cartón como pared puede presentarse como una falsa inconsistencia.

Estas excepciones no deben provocar que se elimine o cambie la pauta, por el contrario deben mantenerse para identificar las inconsistencias verdaderas. Sin embargo el sistema de procesamiento debe ser capaz de manejar las excepciones adecuadamente.

El objetivo de las pautas de consistencia es controlar la calidad de los datos de la investigación. Ante más pautas y más exhaustivas mayor la calidad de los datos y por el contrario, ante menos pautas menor calidad. Entonces, es responsabilidad del investigador principal determinar el grado de calidad deseado y por consecuencia la clase y cantidad de pautas de inconsistencia. Por su parte el informático deberá apoyar esta labor garantizando que la pautas se preparen correctamente.

Es prácticamente imposible una investigación libre de errores. Siempre existirá la posibilidad que en alguna fase del trabajo se incorporen errores en los datos que no sea posible detectarlos. Por ejemplo, una persona puede declarar su peso en 85 kilogramos, siendo al real 82. Para este caso es muy difícil preparar una pauta de consistencia que haga saltar este error y como consecuencia no podrá ser corregido. Sin embargo, este error puede no tener implicaciones con los objetivos de la investigación, ya que por ejemplo, la precisión con que se miden las variables así lo permite o por algunas otras consideraciones estadísticas o demográficas. Por lo tanto, la definición de las pautas de consistencia debe ubicarse dentro de los objetivos y alcances de la investigación y no simplemente para localizar errores como un fin en sí mismo.

Las pautas de consistencia se clasifican en:

pautas de rango: cuando una ocurrencia de una entidad de la base de datos contiene un valor fuera del dominio de la variable. Por ejemplo, el sexo de la persona distinto de 1,2 (hombre y mujer).

pautas de consistencia (propiamente dicha): cuando dos o más variables de la misma o diferente entidad pero del mismo cuestionario tienen valores inconsistentes entre sí. Por ejemplo, inconsistencia entre la edad de la persona y su nivel educativo (menor de 12 años con estudios universitarios).

pautas de estructura: cuando la cantidad de ocurrencias totales (o de cierto tipo) de una entidad es inconsistente por sí misma, o no corresponde o con algún otro atributo del cuestionario. Por ejemplo, dos o más ocurrencias de personas con relación de parentesco "cónyuge" en el mismo hogar, o más personas en la vivienda que lo declarado explícitamente como miembros de la vivienda.

La definición de las pautas de inconsistencia se debe hacer por medio de un conjunto de expresiones condicionales o predicados independientes entre sí, que indica cada uno una condición de error. Esta forma de especificar las pautas se preparó tratando de mantener compatibilidad con el lenguaje del CONCOR [US Bureau of the Census-95A].

Cada pauta de consistencia tiene la forma de una expresión booleana:

$$\mathbf{exp_1\ op\_rel_1\ exp_2\ [op\_log\ exp_3\ op\_rel_2\ exp_4\ ...]}$$

en donde:

exp: es una expresión aritmética que puede estar formada por nombres de variables, operadores aritméticos (+, -, *, /, ^), paréntesis y constantes.

op_rel: es un operador de relación (=, $\neq$, $\langle$, $\rangle$, $\leq$, $\geq$) que permite comparar dos expresiones aritméticas.

op_log: es un operador lógico (Y, O) que permite conectar dos expresiones lógicas. También se puede usar el NO y los paréntesis para construir expresiones más complejas ([ ]: denota opcional).

para permitir una completa especificación de las pautas se acepta:

variables indizadas: en algunos casos es necesario hacer referencia al valor de una variable para una ocurrencia en particular de la entidad. Para esto se puede calificar la variable con un índice numérico que contenga el número de la ocurrencia requerida (ver ejemplo).

especificación de rangos: muchas variables tienen un dominio formado por uno o más rangos por lo que las constantes pueden ser especificadas de la siguiente manera:

[n1-n2, n3-n4,....]

permitiéndose la omisión del guión y del límite superior cuando el rango es un solo número.

constantes especiales: existen algunos valores que pueden pertenecer al dominio de las variables y que tienen un significado especial:

NA: el no aplica es el código adecuado a una variable que no debe tener respuesta en esa ocurrencia.

agregaciones: para especificar las pautas de estructura se puede emplear la siguiente notación:

TOTAL(variable|entidad) [PARA condición]

que sirve para indicar que se trata del total de la suma de la *variable* o el total de casos de la *entidad* para las ocurrencias en que se cumpla la *condición*.

Es conveniente aclarar que la aplicación de las pautas de consistencia se hace a nivel de cuestionario. Esto quiere decir que si la pauta hace referencia a variables de una entidad, esta debe aplicarse a cada ocurrencia de la entidad dentro del cuestionario. Ahora, las variables indizadas hacen referencia únicamente a una ocurrencia específica de la entidad.

ejemplos:

-La pauta para comprobar que el valor del sexo de la persona no pertenece a su dominio es:

sexo ≠ [1-2]

-El último grado o año aprobado en la enseñanza regular es un conjunto de rangos que puede ser especificado de la siguiente forma:

ultimo_grado ≠ [00, 01, 11-16, 21-25, 31-37]

-La siguiente pauta genera una inconsistencia si la vivienda no tiene luz eléctrica (luz=3) y el principal combustible para cocinar es la electricidad (combustible=1).

luz = 3 Y combustible = 1

-En los censos de población y vivienda se acostumbra a enumerar al jefe del hogar como la primera persona, entonces la siguiente pauta da un error si el sexo del jefe es igual al sexo del cónyuge (relación=2).

sexo = sexo(1) Y relación = 2

-En algunas investigaciones se cuestiona si la madre de la persona está presente en la vivienda (madre=1) y si es así cual es su número de secuencia dentro del cuestionario (numero_madre). Entonces es posible comparar la edad de la persona con la de su madre:

madre = 1 Y edad > edad(numero_madre) - 15

-Puede recolectarse información redundante que sirve como control. Un caso es el ingreso total del hogar y el ingreso de los hombres y de las mujeres, además

del ingreso de cada miembro (ingreso). Entonces se puede controlar los totales con:

ingreso_mujeres ≠ TOTAL(ingreso) PARA sexo =2

-Otro chequeo importante puede darse cuando la vivienda es desocupada (6,7,8,9), pues no deben existir ocurrencias de la entidad personas, entonces:

tenencia = [6-9] Y TOTAL(personas) ≠ 0

-Para comprobar que solamente las mujeres de 15 o más años tengan hijos se puede incluir una pauta como:

numero_hijos ≠ NA Y (sexo = 1 O edad < 15)

A continuación se presentan unos comentario adicionales para indicar que las pautas deben ser exhaustivas, no redundantes y no contradictorias.

Es necesario ser exhaustivo en las pautas que se construyen y no se debe dejar condiciones que permitan filtrarse errores que no se detecten. Por ejemplo, si se decide

incluir una pauta que compruebe que cuando la vivienda no tiene servicio sanitario la pregunta sobre su exclusividad debe estar sin respuesta, es necesario también incluir la otra condición que dice que si la exclusividad del servicio sanitario no tiene respuesta, la vivienda no debe tener servicio sanitario.

Por el contrario, es importante no introducir pautas redundantes, o en otras palabras, pautas que no son necesarias pues otra pauta detecta el mismo error en todos los casos.

Las pautas contradictorias son una falla que origina muchos inconvenientes, principalmente a la hora de detectar los errores ya que en todos o en una gran cantidad de cuestionarios saltarán errores, y su origen es una pauta mal definida.

3.3.4 Llaves primarias

Para la localización de los datos de cada cuestionario, es indispensable su identificación precisa. Para esto cada cuestionario debe tener una llave primaria que define de manera única a la entidad raíz.

Para una gran cantidad de investigaciones, esta llave puede estar compuesta por varias partes que identifican las áreas geográficas o institucionales para las cuales se recolecta la información. En un Censo de Población y Vivienda, los cuestionarios están identificados por los códigos (jerárquicos) de las estructura político-administrativa hasta

llegar al área de enumeración o segmento censal. Luego cada cuestionario tendrá además el código de la vivienda a la cual corresponde. Para más información véase el apartado de recolección de datos en esta misma guía.

Las otras entidades de la base de datos dependientes de la raíz tendrán una llave jerárquica compuesta. Por ejemplo, los hogares tendrán la misma identificación de la raíz más el código del hogar, y la persona tendrá la llave del hogar más el número de la persona.

3.3.5 Definiciones incompletas

En algunas ocasiones, como por ejemplo en las encuestas de opinión, no se conoce de antemano las respuestas que se obtendrán para varias preguntas (preguntas abiertas), y será hasta que se recolecte la información y se estudien estas repuestas que se podrá definir el dominio de las variables correspondientes. De tal forma que el modelaje de datos quedará pendiente e incompleto hasta que finalicen otras etapas del procesamiento.

Esta situación presenta algunas variantes. En algunos casos es posible delimitar el máximo número de categorías y así establecer el rango de valores de la variable; en otros casos esto no será factible. Otra posibilidad se da con las preguntas semicerradas que tienen bien definido el conjunto de valores posibles excepto uno que acostumbran a

presentar como "otros", de donde se pueden obtener una o más categorías adicionales para esta variable. En todo caso ver más adelante el apartado sobre preguntas abiertas.

3.4 Diseño del Cuestionario

3.4.1 Consideraciones iniciales

El diseño de un cuestionario es un tarea técnica y multidisciplinaria bien compleja en la cual deben participar principalmente los investigadores que dirigen el proyecto y los informáticos a cargo del procesamiento. Sin embargo, se puede requerir además el apoyo de otro tipo de profesionales como diseñadores gráficos, lingüistas, etc.

A pesar de lo anterior, esta guía pretende tratar solamente los aspectos computacionales que se encuentran en el diseño de cuestionarios.

También es necesario hacer algunas aclaraciones en torno a la terminología. Comúnmente se emplean los términos boleta, cuestionario y formulario para referirse al instrumento de recolección que se utiliza en la investigación.

"La boleta censal puede ser dividida lógicamente en dos partes distintas, cuestionario y formulario. Por cuestionario se entiende la redacción de las preguntas y el establecimiento de las alternativas para sus respectivas respuestas. El formulario es la fórmula con la cual se hacen las preguntas, la disposición de las mismas en el papel, tipo de letra, tamaño, color, etc." [Silva-85]

Sin embargo, ninguna de las definiciones anteriores se acerca al concepto requerido en esta guía, que es ajeno tanto de la propia redacción de las preguntas como de su distribución y formato. Aquí lo importante es el contenido lógico de la boleta, que comprende conceptos como la identificación, definición de unidades de análisis, saltos, etc.

Por lo tanto, se utilizará el término **formulario** para referirse al instrumento físico de recolección de datos y el concepto **cuestionario** para hacer mención a su contenido y estructura lógica.

El diseño de un formulario está influida por la forma en como se recolecta la información. En una mayoría de investigaciones se dispone de un **entrevistador** que formula las preguntas y anota las respuestas, y de un **informante** que responde a la entrevista. Sin embargo, en algunas ocasiones el entrevistador observa directamente (sin informante) el hecho y anota la respuesta, o por el contrario cuando el informante recibe directamente el cuestionario (sin entrevistador) y lo completa y devuelve.

Independientemente de las consideraciones anteriores, que afectan en mayor grado aspectos de forma, el formulario se puede entender como un conjunto de preguntas. Cada pregunta esta formada por:

- texto que cuestiona al informante o guía al entrevistador
- área para anotar la respuesta

Además, es posible tener en el formulario instrucciones específicas sobre el orden de llenado. Más adelante se detallan todos estos aspectos.

3.4.2 Preguntas y variables

No existe una relación uno a uno entre pregunta y variable. Es posible, que de una sola pregunta se obtenga la respuesta para más de una variable y también es posible, que algunas preguntas no colaboren con el conjunto de datos. Por ejemplo, en algunas ocasiones se pregunta el nombre de la persona para, entre otras cosas, facilitar la formulación de otras preguntas, pero no se introduce al computador y no pasa a formar parte del conjunto de datos. Por otro lado, algunas preguntas como el tipo de vivienda, permite obtener el valor a dos variables que son propiamente, el tipo de la vivienda y además la tenencia (propia, alquilada etc.).

Las preguntas que conforman un cuestionario, pero aún más específicamente, las respuestas a estas preguntas deben convertirse en los valores de cada una de las variables o atributos que se producen del modelaje de los datos. De tal forma que el cuestionario es el instrumento para obtener el contenido de todas las variables.

3.4.3 Clasificación de las preguntas

Para efectos del diseño del cuestionario, las preguntas se pueden clasificar en dos grupos. Las preguntas de identificación y las preguntas de contenido.

Las preguntas de identificación en muchos casos no se formula directamente al informante sino que se obtiene de los datos referentes a la organización de la recolección. Por ejemplo los códigos del segmento censal, y el número de la vivienda a entrevistar.

Estas preguntas deben identificar de manera total al cuestionario, en otras palabras, no debe haber otro cuestionario en la investigación que tenga el mismo conjunto de códigos.

Como se describió antes, lógicamente cada cuestionario se representa como una estructura jerárquica de entidades donde cada una puede tener varias ocurrencias, con excepción de la entidad raíz que tiene una única ocurrencia, y cuya llave primaria está formada por las variables correspondientes a las preguntas de identificación del cuestionario.

Por su parte, las preguntas de contenido se refiere propiamente al contenido de la investigación.

3.4.4 Tipos de preguntas

Para los efectos propios del procesamiento de los datos las preguntas se pueden organizar en varios tipos. El criterio para realizar esta tipificación es la forma en que se registra la respuesta, que no es otra cosa más que el propio contenido de la variable que se trata de obtener por medio de esa pregunta. Por tal motivo, el tipo de pregunta y la forma en como se recoge la respuesta se debe obtener de la variable que origina esa pregunta.

Entonces, los tipos de preguntas que se pueden encontrar en estas investigaciones son:

preguntas con respuestas categorizadas: Estas preguntas se refieren a variables que tienen un dominio o conjunto de valores bien definido, en el cual cada código representa una categoría de la variable. En este caso es deseable que se presenten en el cuestionario todos las categorías disponibles con sus respectivos códigos para que el entrevistador seleccione la respuesta adecuada. Esto también recibe el nombre de preguntas precodificadas. Por ejemplo:

Estado Civil:

O_1	Soltero	O_4	Divorciado
O_2	Casado	O_5	Unido
O_3	Viudo	O_6	Separado

Entonces, el usuario marca el círculo correspondiente a la categoría del estado civil de la persona entrevistada. Note que todas las categorías están presentes.

Existe una excepción en este tipo de preguntas, que se da cuando el conjunto de respuestas o categorías es demasiado grande, por lo que no es posible hacer la precodificación. Como ejemplo se tienen variables como la ocupación de la persona, o la causa de muerte, que son conjuntos de códigos bien definidos pero sumamente grandes. En estos casos las preguntas se pueden tratar como preguntas abiertas que luego son codificadas.

preguntas que generan variables continuas: La característica de este tipo de variables es que tienen un dominio de valores continuos dentro de un rango delimitado. En otras palabras, el valor de la variable no representa una categoría específica sino que es un valor numérico continuo. Por ejemplo:

Edad en años cumplidos: ⌊_⌊_⌋

Entonces, el usuario del cuestionario dispone del espacio para anotar una cifra de hasta dos dígitos que representa la edad de la persona.

Nótese que es posible que una variable continua tenga algunas categorías especiales. En el mismo ejemplo de la edad, la variable podría tener dos categorías adicionales: el código 99 que significa *ignorado* y el 98 que podría ser la categoría para las personas de 98 años y más.

Debe quedar claro que la definición de una variable como continua o categorizada no depende de la propia naturaleza de la variable, sino por el contrario del uso que se le quiera dar. Por ejemplo, en determinada investigación la edad podría convertirse en una variable categorizada ya que se requiere únicamente en grandes grupos.

preguntas abiertas: Este tipo de preguntas, bastante utilizadas en las encuestas de opinión, tienen la característica que no se conoce exactamente el conjunto de valores que se tendrán, ya que será hasta que se estudie la información recolectada que se podrá definir con detalle. Es posible que lo único que se pueda determinar con anterioridad es la cantidad de categorías que se permitirán. Por ejemplo:

¿Porqué razón cree usted que .....?

_____

_____ |_|_|

_____

Como se mencionó anteriormente, las variables categorizadas con un gran número de categorías pueden ser tratadas como preguntas abiertas y se pueden presentar de la siguiente forma:

Causa de muerte: _____ |_|_|_|_|

Es posible también, que alguna variable categorizada tenga una respuesta abierta. Por ejemplo, todas las categorías de la pregunta están bien precodificadas pero existe la posibilidad que el informante responda de manera imprevista, entonces se acostumbra a presentar una respuesta adicional como:

"Otros_____ (especifique).

La principal característica de este tipo de variables es que requieren de **codificación**. Esto significa que luego de la recolección de la información, la respuesta a estas preguntas en el cuestionario no tienen el valor que se debe introducir al computador, y que deben pasar por el proceso de codificación, justo antes de la digitación. Este proceso, que se describe en detalle más adelante, consiste en la determinación del valor de la respuesta de acuerdo a lo anotado por el usuario en el cuestionario.

3.4.5 Orden de llenado del cuestionario

3.4.5 Orden de llenado del cuestionario

El orden en que se deben hacer las preguntas, que es el mismo orden en que se dispongan el formulario, está determinado por distintos factores. Algunos de estos son relacionados con el propio contenido de la pregunta y definido por los investigadores y encargados del tema sustantivo. Por otra parte, los otros factores son de importancia únicamente para el posterior procesamiento de los datos.

El orden de las preguntas en un cuestionario está determinado por la estructura propia de las preguntas. Generalmente las preguntas se agrupan en secciones temáticas que están organizadas jerárquicamente. La primera gran división se hace de acuerdo a las entidades comprendidas en la investigación, posteriormente pueden existir otras subdivisiones de acuerdo a los propios temas y a otros subconjuntos de casos que se derivan de la definición de las subclases hecha en el modelaje de datos.

Por ejemplo, las preguntas acerca de la vivienda se agrupan en una sección de la vivienda; las preguntas sobre las personas se agrupa en una sección sobre las personas, y dentro de esta sección se tendrá subsecciones para las preguntas para las mujeres, para la población económicamente activa, etc.

Además, es importante ordenar las preguntas dentro de las secciones de acuerdo a la subclase a la cual pertenezca. Por ejemplo, pueden existir algunas preguntas para toda la población, otras para los mayores de 6 años, otras para los mayores de 12, etc.

De aquí se deriva el concepto de "salto" que se describe en el apartado sobre la digitación. Un salto es la acción que se toma cuando un grupo de preguntas no aplica a determinada ocurrencia de una entidad, entonces, el flujo de llenado del cuestionario salta a otra parte o sección del formulario cuando se cumple la condición correspondiente.

Por ejemplo, cuando se llega a la sección de fecundidad en un censo de población, los hombres y las mujeres menores de 15 años saltarán a la sección siguiente.

Entonces, cada pregunta se hace a un universo determinado de casos que conforman una subclase. Para muchas preguntas, este universo es la totalidad de los casos, sin embargo, para otras preguntas, el universo corresponde a un subconjunto de los casos.

Para el ejemplo anterior, el universo es:

sexo = 2 y edad >= 15.

Entonces, cada pregunta tiene asociada una expresión lógica que define el conjunto de casos para los cuales tiene sentido. Para los casos en los que la expresión lógica es falsa, la respuesta a la pregunta es el valor nulo (NULL).

3.5 Los procedimientos de codificación

En el modelaje de los datos se definió el dominio o conjunto de valores permitidos de cada variable y durante la recolección de la información (en el campo) se obtienen las respuestas por parte del informante, a las preguntas en el cuestionario . Estas respuestas no se dan en términos de los códigos definidos para las variables asociadas a las respuestas, por el contrario, el informante suministra la información en forma de frases, palabras o simplemente un texto que debe ser "traducido" a los códigos aceptados. Por lo tanto, se define la codificación como el proceso mediante el cual se asigna un código válido del dominio de la variable a cada respuesta obtenida en el llenado del cuestionario.

Cada respuesta tiene su propio sistema de codificación que lleva a conformar adecuadamente el dominio de la o las variables producto de la pregunta. Existen distintos tipos de codificación que se adaptan a las características de cada variable. Estos tipos de codificación se pueden dar en distintos momentos del procesamiento de los datos, ya sea en la propia entrevista, en un proceso particular de codificación y hasta en la digitación.

Es posible que el mejor momento para realizar la codificación sea la propia entrevista, ya que se dispone del informante, o del evento que genera la respuesta, para obtener información adicional que mejore la calidad del dato. Sin embargo, esto no siempre es posible debido principalmente a las características propias de cada variable y a la disponibilidad de recursos del entrevistador (capacitación, computador de mano o manuales de códigos, tiempo limitado).

3.5.1 Tipos de codificación

Los tipos de codificación que a continuación que se detallan y que se basan en [NACIONES UNIDAS-91] son la autocodificación y precodificación que se realizan en la propia entrevista, la codificación manual que es la tradicional etapa de codificar en la oficina y por último la codificación automática que se da durante la digitación. También se comenta la posibilidad de utilizar varios tipos de codificación en una misma variable.

Autocodificación

En este caso la respuesta no requiere de codificación o en otras palabras es autocodificada. Esto quiere decir que la propia respuesta emitida por el informante contiene explícitamente el código adecuado. Las variables continuas son el ejemplo más claro de este tipo de codificación. La edad de la persona, el ingreso del hogar o la talla del niño, son variables en las cuales el entrevistador anota exactamente lo que el

informante dice, y esta respuesta es el código válido que la variable tendrá una vez introducida al computador.

Precodificación

Aquí se clasifican las preguntas que llevan a variables con un dominio compuesto de categorías. La idea es incluir en el diseño del formulario las posibles respuestas a la pregunta para que el entrevistador seleccione la adecuada de acuerdo a la respuesta que dio el informante. El texto de cada una de las posibles respuestas debe estar acompañado del código (preferiblemente numérico) que el digitador debe posteriormente introducir al computador. El ejemplo más común puede ser el sexo de las personas 1-Hombre, 2-Mujer, en donde el entrevistador marca la casilla según lo que dice el informante, y posteriormente el digitador introduce al computador el código correspondiente.

Una variante a esta manera de precodificar se da en los formularios de tipo "horizontal" en donde cada ocurrencia (por ejemplo, la persona) ocupa una línea, y cada columna es una pregunta. Entonces en el encabezado de la columna se indican los códigos, y el entrevistador anota el código correspondiente en la casilla.

Codificación manual

Este es el proceso tradicional de codificación en donde los cuestionarios son tratados por personal de "oficina" llamados codificadores que asignan un código válido a cada una de las preguntas que lo requieran. Este tipo de codificación se aplica a las preguntas abiertas que no se conoce como serán sus respuestas, o a las preguntas precodificadas con una cantidad tan grande de códigos que no es posible imprimirlos en el cuestionario.

Entonces, el trabajo de los codificadores es determinar el código adecuado a la respuesta anotada por el entrevistador en el cuestionario. Para realizar esta tarea deben tener un "Manual de Codificación" o equivalente en donde se detalle el procedimiento exacto que se debe seguir para realizar la codificación. Este manual puede contener un detalle de todos los códigos aceptados con su correspondiente descripción. Por ejemplo, La Clasificación Internacional Uniforme de Ocupaciones (CIUO) de la OIT.

Como ejemplo de variables con grandes conjuntos de códigos se puede mencionar la ocupación de la persona, las causas de muerte y la ubicación geográfica de residencia.

Codificación automatizada

Se entiende por codificación automatizada la asignación del código correspondiente a la variable en el momento de la introducción de los datos al computador. En otras palabras, se le confiere al digitador la potestad de determinar cual es el código correspondiente a la respuesta (textual) anotada por el entrevistador. Evidentemente este proceso de

selección de un código se hace con ayuda de la computadora y entra dentro de la categoría de *codificación asistida por computador.*

Codificación mixta

Es posible que para lograr la codificación final de una variable se requieran distintos tipos de codificación. Una pregunta autocodificada puede tener una categoría especial que requiera de codificación manual o automatizada, o también una pregunta con respuestas precodificadas puede tener una categoría abierta que también requiera de otro tipo de codificación.

3.5.2 Codificación asistida por computador

Tanto la codificación manual (en algunos casos) como la automatizada hacen uso de un computador para la selección del código adecuado a la respuesta, lo que introduce este proceso dentro de la temática de la computación conocida como *Computer-Assisted (CA).* En este caso la situación es bastante definida pues se parte de un texto escrito por el entrevistador en el cuestionario, y se debe llegar al código correspondiente. Se identifican las siguientes categorías de codificación asistida por computador:

- el usuario introduce las primeras letras del texto y el computador presenta todas los posibles códigos que concuerden con esas letras. Aquí se requiere de una tabla auxiliar en el computador con una llave primaria con la descripción del código.

- para códigos estructurados es posible formar un conjunto de opciones organizadas jerárquicamente, en dónde el usuario pueda seleccionar alternativas en los distintos niveles hasta completar la definición completa del código adecuado. Por ejemplo, la codificación de los lugares de residencia que están formados por códigos jerárquicos caben en esta categoría.

- el usuario puede digitar totalmente el texto escrito en el cuestionario, y entonces, el computador hace una búsqueda en la base de datos y presenta en la pantalla una lista de los códigos posibles ordenados de mayor a menor de acuerdo al grado de éxito logrado en la búsqueda. Esto requiere de la construcción de sofisticadas bases de datos de recuperación de información textual con archivos invertidos, elaborados índices u otras técnicas.

- los sistemas expertos pueden servir de alternativa para las variables sumamente complejas de codificar, en las cuales el usuario del computador puede entrar en un "diálogo" con el experto en donde se suministre la información *necesaria para* que la máquina emita su opinión sobre el código adecuado.

4. CAPTURA Y PROCESAMIENTO DE DATOS

4.1 Consideraciones generales del proceso

4.1.1 Las etapas del proceso

Aquí se detallan todas las etapas del procesamiento de un censo o encuesta desde la recolección de datos en el campo hasta la obtención de un conjunto de datos consistente. Estas etapas y sus características se muestran en la figura 4.1 y requieren ejecutarse en el orden descrito, y probablemente están presentes en todas las investigaciones.

4.1.2 Procesamiento por lotes

La ejecución de un censo o encuesta de alguna magnitud requiere generalmente el manejo de una gran cantidad de documentos o cuestionarios que se mueven a través de las distintas etapas del procesamiento. Esto hace que se requiera organizar las labores de campo y oficina de tal manera que este funcionamiento se realice en orden y con un control adecuado.

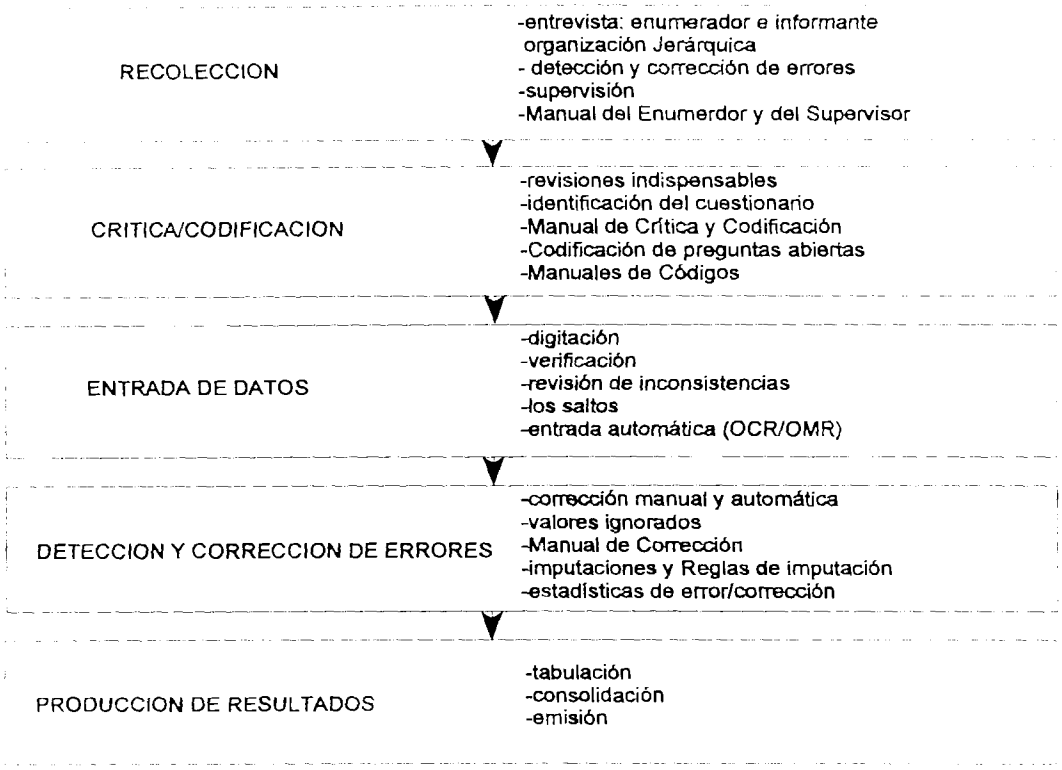


Figura 4.1: Características de las etapas del procesamiento

Usualmente, este manejo de documentos debe ser organizado como un sistema de producción por lotes. Estos lotes deben corresponder a las áreas de enumeración o a la división del trabajo de recolección que se utiliza en la investigación.

Esto quiere decir que la información se recoge para un área determinada, entonces, el conjunto de boletas de cada área pasa a conformar un lote que permanecerá junto durante el resto de las etapas del procesamiento.

Esta forma de organizar el trabajo y que se conoce con el nombre de Control Primario tiene grandes beneficios ya que permite dentro de otras cosas, controlar los documentos y asignar las cargas de trabajo de manera eficiente.

Este sistema de manejo de documentos puede implementarse de manera manual con unas cuantas funciones extras al personal encargado, o por el contrario puede informatizarse para controlar el flujo del trabajo por medio del computador.

Sea cual sea la forma de funcionar el sistema debe tener las siguientes lineamentos:

- el lote debe identificarse de manera única, preferiblemente con la identificación del área de enumeración a la cual pertenece.

- cada lote de documentos debe pasar una y sólo una vez por cada proceso, evitando así la duplicación de trabajo (y datos) y la posible omisión.

- comparando las identificaciones de los lotes procesados con las listas de áreas de enumeración se puede controlar la cobertura de la investigación.

- el rendimiento de los funcionarios (críticos-codificadores-digitadores-correctores) debe adecuarse a ciertas normas que serán controladas por la cantidad de lotes

procesados por unidad de tiempo, por lo que cada lote será manejado por una sola persona en cada proceso.

- la evaluación de los rendimientos de los procesos servirá para "balancear la línea de producción" modificando el número de personas en cada proceso o mejorando las condiciones de trabajo para lograr cumplir con los planes.

- para establecer los controles de calidad se pueden muestrear lotes en determinado proceso para realizar un revisión adicional.

- el orden de procesamiento de los lotes debe coincidir con las políticas de prioridades establecidas.

4.1.3 Procesamiento descentralizado

La característica de poder realizar los procesamientos en lotes de cuestionarios permite también pensar en procesar la información en distintas ubicaciones físicas, probablemente en oficinas regionales o departamentales que reciban los cuestionarios de las áreas de enumeración que dependan de ellas.

De esta forma, los procesos de crítica/codificación, digitación y aún la depuración de datos pueden realizarse en oficinas descentralizadas, y luego buscar la consolidación de los datos a niveles centrales.

Esta modalidad de trabajo es posiblemente útil en grandes investigaciones en las cuales sea factible esta descentralización, y presenta varias ventajas como mantener los procesos cerca del lugar de recolección, el trasiego de documentos se realiza localmente, y a niveles centrales llega únicamente la información en medios legibles por el computador. Sin embargo, la decisión de la centralización o descentralización está determinada probablemente por el volumen de la investigación y los recursos disponibles.

4.2 Recolección de datos

Esta guía hace referencia en muchas ocasiones a investigaciones de tipo poblacional, que son probablemente las más comunes; sin embargo, la gran mayoría de los temas tratados aquí son aplicables también a una buena cantidad de estudios que requieren recoger y procesar datos y luego presentar los resultados.

De la misma manera, la terminología empleada para la recolección de los datos se refiere a un enumerador que visita una vivienda y anota en un cuestionario las respuestas que brinda un informante. Esto no es regla general ya que en muchas ocasiones no existe el informante ya que es el propio enumerador que obtiene la información, o en el caso

contrario también es posible la no participación de un enumerador y que el informante complete personalmente el cuestionario.

En términos generales se puede decir que el proceso de recolección se basa en el registro de los datos producidos por la ocurrencia de un determinado evento. En el caso de la investigación de la población, el evento puede ser la existencia de un grupo de personas que habita una vivienda y que tiene ciertas características en un momento dado. En otro tipo de estudios este evento puede ser, por ejemplo, una muestra de aguas contaminadas de un río en cierta fecha y lugar, el ingreso de un expediente médico de un enfermo de cáncer, etc.

4.2.1 Organización de la Recolección

Generalmente los estructuras de recolección de los datos son de forma jerárquica; ya sea si se trata de una investigación territorial, en donde las unidades geográficas son divididas en áreas de enumeración, o también cuando se utiliza una determinada estructura institucional en sus distintos niveles de desagregación.

Esta característica permite que la recolección se organice de la misma forma, estableciendo mecanismos de enumeración, revisión y consolidación hasta que los cuestionarios lleguen a niveles regionales o centrales.

En las investigaciones de cierta magnitud, se requiere de equipos de enumeradores que cubren determinada área (territorial o institucional) y que son dirigidas por un supervisor que controla su rendimiento y efectividad. En estudios más pequeños, es probable que la supervisión se realice únicamente a nivel central y con pocos enumeradores que se mueven a través de las área.

4.2.2 Detección y Corrección de Errores

Se dice que una investigación es tan buena como la calidad del proceso de la recolección de los datos, por lo que generalmente se hacen grandes esfuerzos en la planeación, organización y ejecución de esta etapa.

La gran mayoría de actividades que se pueden realizar para este fin están fuera de los alcances de esta guía, sin embargo existen algunas consideraciones trascendentales relacionadas con el manejo de los errores.

Es sumamente importante incorporar funciones de detección y corrección de errores en la etapa de recolección que permitan mejorar la calidad, ya que aún se dispone del evento que generó el dato y es posible determinar con mucha precisión el valor correcto de la variable con error. Lamentablemente, esta es una actividad que requiere bastantes recursos extras (tiempo y personas) y en la mayoría de los casos debe ser limitado a comprobación de pautas muy importantes.

Estas funciones de detección y corrección pueden ser aplicadas en la propia enumeración o entrevista y en la supervisión (si existe en el campo), y sobresalen las siguientes:

- llenado de todas las preguntas. Al final de la entrevista se puede revisar rápidamente que todas las preguntas (que se espere) tengan una respuesta. De esta manera se minimiza el problema de la información faltante.

- datos para todas las ocurrencias. Para cuestionarios con varias entidades, es posible comprobar que la declaración del número de ocurrencias concuerde con la existencia de datos para cada ocurrencia. Por ejemplo, si se declaró la residencia de 5 personas en el hogar debe existir información para cada una de las 5 personas.

- totales de control. En algunas ocasiones se preguntan cifras parciales y totales lo que permite comprobar que el total es la suma de los parciales. Por ejemplo, se puede preguntar sobre la cantidad de personas de sexo masculino, de sexo femenino y el total general.

- variables muy importantes. Por ejemplo, en una investigación poblacional se puede instruir al enumerador para que, al momento de llenar el cuestionario, compruebe que el sexo declarado concuerda con el nombre de la persona.

A nivel de supervisión es posible (y a veces indispensable) incorporar además algunas revisiones importantes referentes principalmente a la cobertura de la investigación. En este tipo se encuentran:

- omisión de casos. De acuerdo a lo programado, es recomendable saber si se obtuvo el cuestionario correspondiente para cada una de las observaciones que se debían hacer.

- duplicación de casos. De la misma forma anterior, se puede comprobar que no existan dobles cuestionarios para la misma observación.

- revisión de una muestra. Es común que entre las funciones de la supervisión se incluya la revisión (un tanto rigurosa) de una pequeña muestra de los cuestionarios de cada enumerador, de tal forma que se determine la calidad del llenado de los formularios y la posible detección de errores sistemáticos.

Las revisiones de inconsistencias que se decidan incorporar en la fase de recolección deben estar integradas con la *Estrategia Global de Manejo de Errores*. De esta forma será posible obtener una mejor calidad en los datos al final de todo el proceso. Por ejemplo, si se planea utilizar determinada variable para controlar las imputaciones automáticas a otras, entonces es aconsejable incorporar revisiones a esta variable en las primeras fases del proceso, como en la recolección de los datos, de esta forma se

garantiza que la variable que se empleará de base para la imputación tiene una alta calidad.

Todas estas revisiones deben estar claramente detalladas en los documentos que instruyen los funcionarios encargados de la recolección. Estos documentos son el Manual del Enumerador y el Manual del Supervisor.

4.3 Crítica/Codificación

En este proceso se agrupan las actividades de "oficina" que se realizan con los cuestionarios antes de la entrada de los datos al computador.

La crítica y la codificación pueden verse como una sola actividad realizada por la misma persona o como dos actividades independientes y consecutivas, cada una tiene sus características y objetivos. Generalmente, y lo recomendable si es posible, es la realización de las dos actividades por la misma persona en un solo proceso, pero esto depende de los volúmenes de trabajo, el grado de especialización y otros asuntos menores.

Para realizar este proceso, el funcionario encargado debe contar con un "Manual de Crítica/Codificación" en donde se detallen todas las instrucciones del caso. Por otra parte, el resultado de este proceso debe ser un cuestionario perfectamente listo para pasar a la etapa de entrada de datos.

4.3.1 La crítica

La crítica o revisión es el proceso manual en el cual se revisa que cada cuestionario cumpla con algunas condiciones mínimas para que pueda ser introducido al computador.

Esta revisión manual puede ser exhaustiva y rigurosa logrando como producto una altísima calidad en los datos del cuestionario. Sin embargo, por razones de volumen esto no es posible en la mayoría de investigaciones, por lo que la revisión debe limitarse a los puntos de mayor trascendencia, algunos de los cuales se comentan a continuación:

- Identificación del cuestionario. Una revisión indispensable que debe realizarse en casi todos los procesos de crítica, es la verificación de la identificación del cuestionario, ya que será la llave primaria y única forma de localizar el cuestionarios dentro de la base de datos. Si se trabaja por lotes de cuestionarios, la identificación tendrá una raíz en común para todos los casos, cambiando únicamente la parte propia del cuestionario. Por ejemplo, en un Censo de Población y Vivienda, la identificación del lote comprende las variables de la

ubicación geográfica del área de enumeración, y cada cuestionario está identificado adicionalmente con el código de la vivienda correspondiente.

- Cuestionario rechazados. Existen algunos casos graves en que el cuestionario esta totalmente en blanco o con muy poca información o de muy mala calidad, para los cuales se decide no incluirlos en la investigación. Por lo tanto le queda a la crítica la función de excluir este cuestionario del lote.

- Respuestas múltiples. En algunas ocasiones se marcan incorrectamente varias respuestas a una pregunta de respuesta simple. Esto puede ser revisado por el crítico para evitar que el digitador se encuentre ante la imposibilidad de no saber cual código introducir al computador.

- Variables importantes. En todas las investigaciones siempre hay algunas variables especialmente importantes y para las cuales los directores del proyecto esperan una particular calidad. Estas variables pueden ser revisadas durante la crítica aplicando las pautas de consistencia en un proceso de detección manual de inconsistencias. Para la corrección se debe seguir las consideraciones descritas más adelante en el apartado de corrección manual de errores.

4.3.2 La codificación

Como se mencionó en el apartado sobre los procedimientos de codificación, en algunos casos es necesario efectuar una tarea manual para asignar a una respuesta el código que le corresponde. Este es el proceso de codificación manual que se describe aquí, y que consiste en el análisis por parte del funcionario encargado de la información en el cuestionario, y que con la ayuda de un instructivo de codificación, asigne el código adecuado a cada una de las preguntas que lo requieran.

Esta fase puede comprender actividades muy simples como también otras bastante complejas que requieren de personal muy capacitado.

Las preguntas que requieren codificación son todas aquellas para las cuales no es posible codificar la respuesta durante la recolección, y es claro que se conoce desde el diseño de la investigación cuales serán. Por lo tanto, dentro de este diseño debe definirse claramente las preguntas sujetas a la codificación manual y el procedimiento correspondiente en cada caso.

A continuación se comentan los tipos de pregunta que generalmente requieren codificación:

- grandes conjuntos de códigos. Aquí se encuentran una serie de preguntas en las que el conjunto de códigos del dominio de la variable es sumamente grande, por lo que no es posible que el enumerador determine el adecuado. Los ejemplos más

comunes son la ocupación de la persona, la rama de actividad de la empresa en donde labora y el lugar de residencia. Para realizar esta codificación, el funcionario debe buscar en el manual de códigos el texto que anotó el enumerador, y registrarlo en el cuestionario.

- variables especializadas. Cierto tipo de variables, aún con un dominio pequeño, requieren de un codificador especializado. Por ejemplo, para codificar la localización e histología de un tumor es necesario analizar informes de patología y diagnósticos clínicos que probablemente es conveniente que lo realice un médico especialista.

- preguntas abiertas. La codificación aquí requiere del estudio de las respuestas en todos los cuestionarios para determinar el dominio de la variable, y luego efectuar la codificación. En esta categoría se encuentran las preguntas de opinión pública y algunas preguntas con una categoría abierta (otros ...).

Como se mencionó anteriormente, este proceso de codificación es una fuente potencial de errores. El codificador puede asignar un código de manera incorrecta que haga que una pauta de consistencia resulte positiva. Por lo tanto es conveniente que se establezcan los mecanismos para controlar la calidad del proceso. Por ejemplo, el supervisor de los codificadores puede revisar algunos cuestionarios de cada uno de sus subalternos, o cada codificador puede revisar algunos cuestionarios de otro codificador.

garantiza que la variable que se empleará de base para la imputación tiene una alta calidad.

Todas estas revisiones deben estar claramente detalladas en los documentos que instruyen los funcionarios encargados de la recolección. Estos documentos son el Manual del Enumerador y el Manual del Supervisor.

4.3 Crítica/Codificación

En este proceso se agrupan las actividades de "oficina" que se realizan con los cuestionarios antes de la entrada de los datos al computador.

La crítica y la codificación pueden verse como una sola actividad realizada por la misma persona o como dos actividades independientes y consecutivas, cada una tiene sus características y objetivos. Generalmente, y lo recomendable si es posible, es la realización de las dos actividades por la misma persona en un solo proceso, pero esto depende de los volúmenes de trabajo, el grado de especialización y otros asuntos menores.

Para realizar este proceso, el funcionario encargado debe contar con un "Manual de Crítica/Codificación" en donde se detallen todas las instrucciones del caso. Por otra parte, el resultado de este proceso debe ser un cuestionario perfectamente listo para pasar a la etapa de entrada de datos.

4.3.1 La crítica

La crítica o revisión es el proceso manual en el cual se revisa que cada cuestionario cumpla con algunas condiciones mínimas para que pueda ser introducido al computador.

Esta revisión manual puede ser exhaustiva y rigurosa logrando como producto una altísima calidad en los datos del cuestionario. Sin embargo, por razones de volumen esto no es posible en la mayoría de investigaciones, por lo que la revisión debe limitarse a los puntos de mayor trascendencia, algunos de los cuales se comentan a continuación:

- Identificación del cuestionario. Una revisión indispensable que debe realizarse en casi todos los procesos de crítica, es la verificación de la identificación del cuestionario, ya que será la llave primaria y única forma de localizar el cuestionarios dentro de la base de datos. Si se trabaja por lotes de cuestionarios, la identificación tendrá una raíz en común para todos los casos, cambiando únicamente la parte propia del cuestionario. Por ejemplo, en un Censo de Población y Vivienda, la identificación del lote comprende las variables de la

ubicación geográfica del área de enumeración, y cada cuestionario está identificado adicionalmente con el código de la vivienda correspondiente.

- Cuestionario rechazados. Existen algunos casos graves en que el cuestionario esta totalmente en blanco o con muy poca información o de muy mala calidad, para los cuales se decide no incluirlos en la investigación. Por lo tanto le queda a la crítica la función de excluir este cuestionario del lote.

- Respuestas múltiples. En algunas ocasiones se marcan incorrectamente varias respuestas a una pregunta de respuesta simple. Esto puede ser revisado por el crítico para evitar que el digitador se encuentre ante la imposibilidad de no saber cual código introducir al computador.

- Variables importantes. En todas las investigaciones siempre hay algunas variables especialmente importantes y para las cuales los directores del proyecto esperan una particular calidad. Estas variables pueden ser revisadas durante la crítica aplicando las pautas de consistencia en un proceso de detección manual de inconsistencias. Para la corrección se debe seguir las consideraciones descritas más adelante en el apartado de corrección manual de errores.

4.3.2 La codificación

Como se mencionó en el apartado sobre los procedimientos de codificación, en algunos casos es necesario efectuar una tarea manual para asignar a una respuesta el código que le corresponde. Este es el proceso de codificación manual que se describe aquí, y que consiste en el análisis por parte del funcionario encargado de la información en el cuestionario, y que con la ayuda de un instructivo de codificación, asigne el código adecuado a cada una de las preguntas que lo requieran.

Esta fase puede comprender actividades muy simples como también otras bastante complejas que requieren de personal muy capacitado.

Las preguntas que requieren codificación son todas aquellas para las cuales no es posible codificar la respuesta durante la recolección, y es claro que se conoce desde el diseño de la investigación cuales serán. Por lo tanto, dentro de este diseño debe definirse claramente las preguntas sujetas a la codificación manual y el procedimiento correspondiente en cada caso.

A continuación se comentan los tipos de pregunta que generalmente requieren codificación:

- grandes conjuntos de códigos. Aquí se encuentran una serie de preguntas en las que el conjunto de códigos del dominio de la variable es sumamente grande, por lo que no es posible que el enumerador determine el adecuado. Los ejemplos más

comunes son la ocupación de la persona, la rama de actividad de la empresa en donde labora y el lugar de residencia. Para realizar esta codificación, el funcionario debe buscar en el manual de códigos el texto que anotó el enumerador, y registrarlo en el cuestionario.

- variables especializadas. Cierto tipo de variables, aún con un dominio pequeño, requieren de un codificador especializado. Por ejemplo, para codificar la localización e histología de un tumor es necesario analizar informes de patología y diagnósticos clínicos que probablemente es conveniente que lo realice un médico especialista.

- preguntas abiertas. La codificación aquí requiere del estudio de las respuestas en todos los cuestionarios para determinar el dominio de la variable, y luego efectuar la codificación. En esta categoría se encuentran las preguntas de opinión pública y algunas preguntas con una categoría abierta (otros ...).

Como se mencionó anteriormente, este proceso de codificación es una fuente potencial de errores. El codificador puede asignar un código de manera incorrecta que haga que una pauta de consistencia resulte positiva. Por lo tanto es conveniente que se establezcan los mecanismos para controlar la calidad del proceso. Por ejemplo, el supervisor de los codificadores puede revisar algunos cuestionarios de cada uno de sus subalternos, o cada codificador puede revisar algunos cuestionarios de otro codificador.

4.4 Entrada de Datos

4.4.1 La digitación

La digitación es el proceso en el cual se transcriben los códigos desde el cuestionario a un medio legible por el computador. En este proceso interviene el digitador, el computador (o terminal) y el programa de entrada de datos.

La digitación de datos de encuestas o censos tiene sus características particulares y no se puede considerar igual a cualquier captura de datos de otro tipo de sistemas. Dentro de estas características se encuentran las siguientes:

- generalmente el conjunto de datos es muy grande, por ejemplo en la mayoría de los censos de pobiación y vivienda se manejan millones de casos

- se realiza en otro lugar y momento del evento que generó los datos.

- es posible aceptar datos con error.

El proceso de digitación de los datos debe adecuarse a las características de la investigación y a la estrategia global de manejo de errores, y el problema básico es

4.4 Entrada de Datos

4.4.1 La digitación

La digitación es el proceso en el cual se transcriben los códigos desde el cuestionario a un medio legible por el computador. En este proceso interviene el digitador, el computador (o terminal) y el programa de entrada de datos.

La digitación de datos de encuestas o censos tiene sus características particulares y no se puede considerar igual a cualquier captura de datos de otro tipo de sistemas. Dentro de estas características se encuentran las siguientes:

- generalmente el conjunto de datos es muy grande, por ejemplo en la mayoría de los censos de pobiación y vivienda se manejan millones de casos

- se realiza en otro lugar y momento del evento que generó los datos.

- es posible aceptar datos con error.

El proceso de digitación de los datos debe adecuarse a las características de la investigación y a la estrategia global de manejo de errores, y el problema básico es

definir el nivel de revisión de inconsistencias y la forma en que se harán las correcciones.

El programa que se utiliza para la captura de datos puede limitarse a aceptar cualquier dato que el digitador introduzca y a registrarlo en el computador o por el contrario puede aplicar rigurosamente todas las pautas de consistencia impidiendo que entren datos con errores. De una a otra alternativa existe un "continuo" del cual se debe elegir la opción más adecuada para la investigación. En la figura 4.2 se presentan ambos extremos con sus beneficios y desventajas.

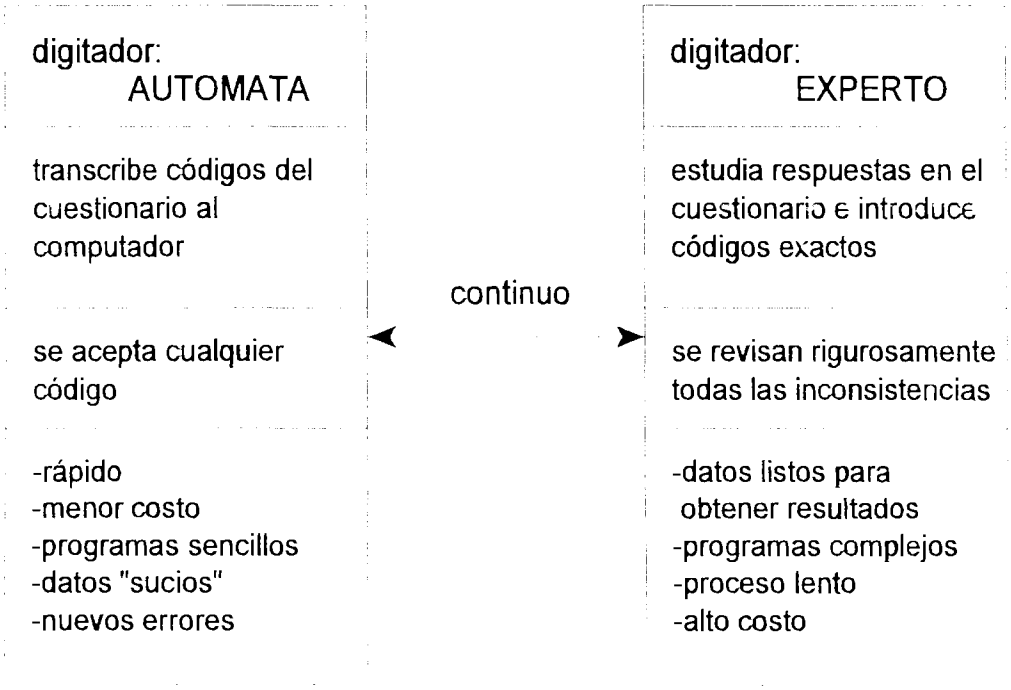


Figura 4.2: Extremos en la definición de la digitación

En la primera posibilidad, en donde se acepta libremente cualquier dato se puede pretender la participación de digitadores con muy poca capacitación ya que serán entrenados únicamente para transcribir, como autómatas, códigos del cuestionario al computador. Debido a lo anterior y a que el programa de entrada de datos es elemental, se presume que se tendrá un costo menor que en otras alternativas. Este proceso de entrada de datos es también el más rápido, ya que el computador prácticamente no "opondrá resistencia" a que el digitador introduzca cualquier código, y éste podrá hacer su trabajo sin interrupciones.

Evidentemente, la desventaja de esta alternativa es en relación a la calidad de los datos. Debido a que no se controla pauta de consistencia alguna, no se podrá detectar los errores en los datos; y más grave aún, los errores producidos en esta misma etapa de digitación aumentarán el conjunto total de inconsistencias y que deberán ser detectadas y corregidas en la siguiente etapa del procesamiento.

En el otro extremo se tiene el proceso de entrada de datos gobernado por un programa que aplica todas las inconsistencias posibles. En este caso se requiere de digitadores sumamente calificados o expertos (probablemente los mismos investigadores) que se enfrenten con un programa que revisa exhaustivamente todas las inconsistencias, debiendo analizar la información en el cuestionario para corregir cada error.

Esta alternativa requiere de muchos más recursos que las otras. No sólo en lo que se refiere a la capacitación de los digitadores, sino también en el tiempo requerido para la introducción de los datos correctos, además del costo de desarrollar programas de entrada de datos complejos. Como beneficio principal se tiene que cuando se termine la digitación (si algún día se logra) se obtendrá un conjunto de datos perfectamente consistente y listo para producir resultados.

Entonces, cada investigación tendrá su adecuado punto en este continuo y el éxito en la digitación de los datos dependerá de que el nivel de aplicación de inconsistencias, y de la forma de hacer la corrección, concuerde con las características de la investigación.

Un punto medio que probablemente sea el ideal para muchas investigaciones tendría características como las siguientes. El programa revisa errores de rango y algunas pocas inconsistencias. Si se detecta alguna error el digitador podrá revisar si se trata de un error de digitación para corregirlo, si no es así, continúa dejando el valor incorrecto. En otras palabras, el digitador no tiene autorizaciór de corregir más que sus propios errores. Todo cuestionario entra a la base de datos, ninguno se rechaza.

4.4.2.1 La verificación

Una de las formas de evitar la introducción de errores en la propia etapa de digitación es realizando una doble digitación, lo que es comúnmente llamado verificación. Este

proceso consiste en volver a digitar la información detectando los errores de la primera digitación; entonces cuando la persona que realiza la verificación detecta una diferencia con lo digitado anteriormente puede observar el cuestionario para determinar si el error es en la verificación o en la propia digitación, en cuyo caso debe corregir el dato almacenado. Debido a que la probabilidad de que el verificador cometa el mismo error que el digitador es muy baja, se espera que la cantidad de errores producto de esta etapa sea mínima.

Se recomienda que la persona que realiza la digitación no sea la misma que digitó la información. Esto porque si se cometen errores sistemáticos o de concepto en la digitación es probable que se cometan también en la verificación, perdiendo así la posibilidad de detectarlos.

La verificación, a pesar de colaborar con que el proceso de entrada de datos sea de alta calidad, provoca un aumento considerable en los costos y en el tiempo total de procesamiento. Los costos se ven afectados tanto por la necesidad de disponer de más digitadores y computadores como también por la elaboración de programas (o facilidades) adicionales para efectuar la verificación.

Verificación parcial

Para disminuir el efecto anterior, es posible limitar la verificación de los datos, tanto a un conjunto de determinadas variables como también a un subuniverso de los casos.

No todas las variables tienen la misma importancia dentro de una investigación. Algunas variables son trascendentales, otros no lo son tanto, por lo que es posible seleccionar un conjunto de variables a las cuales se les hará la verificación.

La verificación puede tener un enfoque de "control de calidad", en donde se obtenga una muestra de los cuestionarios digitados para realizar la verificación. De este modo, si la cantidad de errores de digitación es más alta que determinado parámetro el lote de cuestionarios es rechazado y podrá sufrir una verificación total u otra medida correctiva. El tamaño de la muestra puede variar con el tiempo, ya que al principio debe ser suficientemente grande, pero luego de que los digitadores obtengan experiencia en su labor, puede disminuir considerablemente.

4.4.2.2 Las características de los programas de entrada de datos

Los programas que se desarrollen para la entrada de datos de los censos y encuestas deben presentar una serie de consideraciones particulares:

la interfase

Debe prevalecer la eficiencia y rapidez del proceso sobre la "amigabilidad" y el lujo de la interfase del programa con el usuario. Es importante que la velocidad de digitación esté determinada por la rapidez del digitador y no por el tiempo de respuesta del computador al pasar de campo a campo o de pantalla a pantalla.

La organización de las pantallas y de los datos en su interior, deben acercarse lo más posible a la distribución de las preguntas en el cuestionario.

Es probable que lo más conveniente es que el digitador utilice únicamente su mano derecha para introducir los datos en la parte numérica del teclado, quedando la mano izquierda para manipular los cuestionarios. De tal modo que el uso de otras teclas debe quedar solamente para casos especiales.

El digitador estará observando principalmente el cuestionario y rara vez la pantalla por lo que la mejor forma de emitir mensajes es por medio de un sonido que llame su atención.

la lógica

La unidad de trabajo es el cuestionario por lo que el ciclo de captura de datos y registro en el medio de almacenamiento debe darse para cada cuestionario. Sin embargo, si se

trabaja por lotes de cuestionario, también se tendrá un ciclo para cada lote, digitando la identificación del lote (que será parte de la identificación del cuestionario) una única vez.

A pesar de que la función principal del programa es el de agregar cuestionarios, también se deben tener opciones de consultar, eliminar y modificar los datos

El programa podrá tener dos modalidades de operación que son la digitación y la verificación.

el manejo de los errores

La aplicación de las pautas de inconsistencia puede darse en el momento propio de la captura de cada campo (rango e inconsistencias) o al final de capturar el cuestionario completo (errores de estructura).

Cualquiera que sea la alternativa de chequeo de errores en la captura de datos, el programa deberá reportarlos al digitador y esperar la acción correspondiente. Esta acción dependerá del origen del error, ya que si se trata de un error de digitación el operador ingresa el dato correcto y se elimina la condición de error. Por el contrario, si el origen del error es en una etapa anterior, el digitador tendrá la posibilidad (de acuerdo al diseño del manejo de los errores) de forzar la entrada del dato erróneo para que sea

corregido en etapas posteriores, o por el contrario, podrá rechazar el cuestionario para que sea revisado y corregido manualmente.

los saltos en la secuencia de captura

Como se ha mencionado anteriormente, en muchas investigaciones se tienen grupos de preguntas que se hacen únicamente para un subconjunto de casos. Esta situación es conveniente trasladarla a la digitación para evitar que el operador tenga que teclear en exceso ya que se sabe que determinado grupo de preguntas no será necesario transcribirlas al computador.

Por ejemplo, si se supone que las preguntas de empleo se hacen únicamente a las personas mayores de 12 años no es conveniente que el operador pase una por una todas estas preguntas indicándole al computador que esas variables tendrán un valor nulo.

Los saltos se pueden manejar de manera manual o automática. En la modalidad manual, el operador dispone de una tecla que presiona cuando va a entrar a un grupo de preguntas que desea saltar, entonces el computador se posiciona en el primer campo después de ese grupo y asigna el valor nulo a cada una de las variables correspondientes.

Por el contrario, en la automática es el computador el que decide al salir de un campo, si hace a un salto de acuerdo a los valores de las variables que se han introducido. En esta modalidad se tiene la ventaja que el operador no tiene que preocuparse por los saltos, sin embargo, presenta la desventaja de que si la variable que decide el salto tiene un valor erróneo producirá un salto que no debe darse. Por ejemplo, si la edad se anotó 10 años en lugar de 20, el computador no captará las preguntas de empleo a pesar de que en el cuestionario vienen las respuestas (como debe ser).

La figura 4.3 muestra un esquema del programa de entrada de datos. Nótese que las pautas de inconsistencia pueden estar "empotradas" en el código fuente del programa o por el contrario pueden residir en una tabla externa como en un programa de uso general.

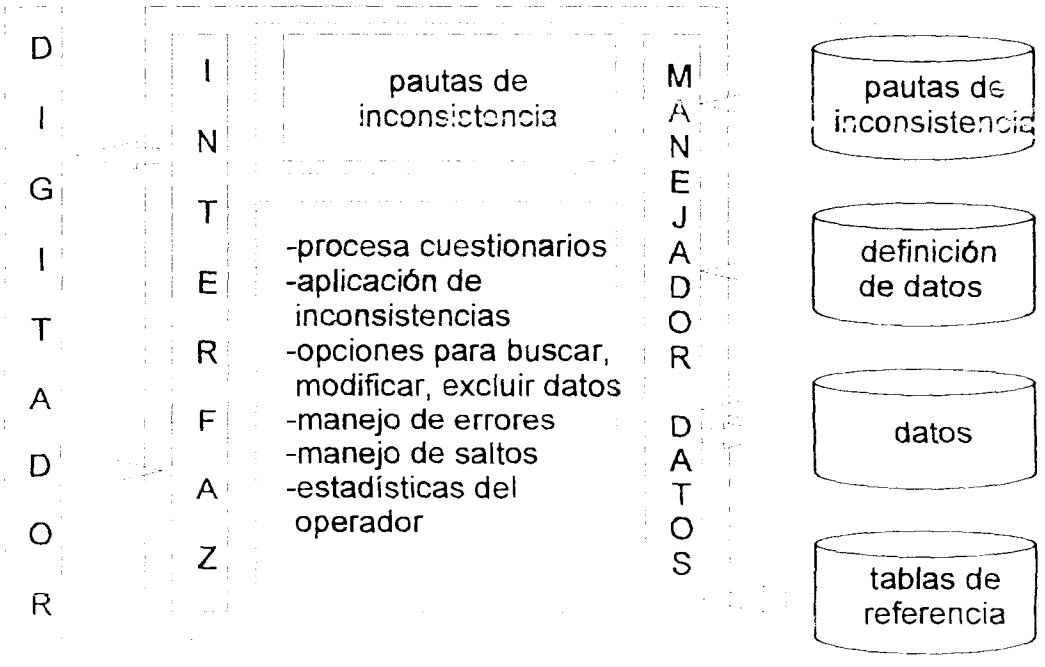


Figura 4.3: Características del programa de entrada de datos

Otra consideración importante es la necesidad de utilizar tablas auxiliares de referencia. Estas tablas contienen conjuntos de códigos como, por ejemplo, la identificación del cuestionario o para variables con un dominio muy grande como la ocupación o la causa de muerte.

4.4.2 Entrada automática

4.4.2.1 Comentario General

Tradicionalmente la digitación ha sido el cuello de botella en el procesamiento de censos y encuestas, y en muchas ocasiones es necesario contratar una gran cantidad de personal que labora en varios turnos, día y noche, para lograr introducir al computador la información de todos los cuestionarios en el tiempo establecido.

Unido a lo anterior, se debe considerar a la digitación como una de las fuentes principales de error, en parte debido a que cualquier actividad con la participación humana esta sujeta a la posibilidad de la existencia de errores.

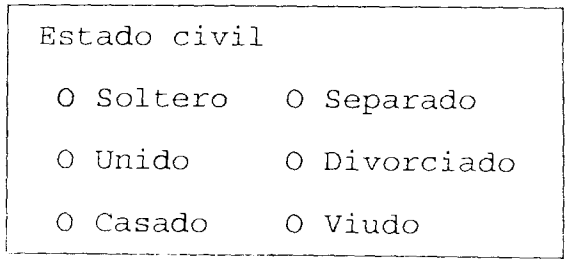
Debido a lo anterior, se han desarrollado equipos y programas que permitan la captura de los datos de manera automática, con el objetivo de minimizar la participación humana, ganando tiempo y calidad en el procesamiento de la investigación.

Estos sistemas de entrada automática de datos comprenden las técnicas como la lectura de marcas ópticas, OMR (Optical Mark Reader) y la lectura de caracteres ópticos, OCR (Optical Character Reader), cada una con su características propias y ventajas y desventajas.

Para todos los casos, el uso de estas técnicas requiere de diseñar gran parte del proceso de la investigación de manera adecuada a la entrada de datos automática. El diseño del cuestionario, el manejo de los errores son los temas más afectados como se comenta posteriormente.

4.4.2.2 OMR: Lectura de marcas ópticas

Estos sistemas utilizan un formulario especialmente diseñado para su uso, en el cual cada pregunta ocupa una área física determinada (generalmente grandes), en donde se presentan todas las posibles respuestas, para que el enumerador durante la recolección "llene" con un lápiz el círculo de la categoría seleccionada. Por ejemplo:



Entonces, el equipo OMR "lee" el área asignada a la pregunta y determina en cual parte del área se encuentra la marca y de esta forma conoce cual es el código de la variable. Los formularios que no pueden leerse se depositan en una bandeja especial para que los funcionarios encargados arreglen el problema.

Esta técnica es especialmente útil para preguntas con respuestas categorizadas como la del ejemplo anterior. Sin embargo, las variables cuantitativas podrán tratarse como categorizadas definiendo una gran cantidad de rangos para presentarlas en el cuestionario de la forma mostrada. Otra alternativa es una distribución tabular donde habrá 10 filas (del 0 al 9) y tantas columnas como dígitos tenga la variable, por ejemplo, el registro de la edad se puede hacer de la siguiente manera:

Edad:	0:	○	○
	1:	○	○
	2:	○	○
	3:	●	○
	4:	○	○
	5:	○	●
	6:	○	○
	7:	○	○
	8:	○	○
	9:	○	○

lo que significa 35 años. Para otras variables com más dígitos simplemente se agregan columnas.

4.4.2.3 OCR: Lectura óptica de caracteres

Estos equipos obtienen un copia gráfica del formulario por medio de un proceso conocido como "scanning" y luego traducen el contenido para obtener el valor de la respuesta de cada pregunta.

Esta técnicas permiten procesar cualquier tipo de variable, pero presentan el inconveniente de que no todos los enumeradores escriben los números de la misma manera, dificultando el reconocimiento de los caracteres.

Para atender los casos en los cuales el OCR no es capaz de entender el caracter, se requiere de personal que utilizando computadores resuelvan el problema tomando la decisión adecuada.

4.4.2.4 Ventajas y desventajas

El uso de lectoras ópticas presenta grandes ventajas en el procesamiento de censos y grandes encuestas como:

- el tiempo requerido para la introducción de los datos al computador es mínimo.

- se elimina la digitación que es una gran fuente de error.

Sin embargo, tienen una serie de problemas que deben ser considerados para evaluar la posibilidad de utilizarla:

- el diseño del formulario se hace con el fin de ser leído por estos equipos y no necesariamente para mejorar la calidad de la recolección.

- los costos son altos, ya que los equipos no podrán ser redistribuidos a otras actividades luego de la captura. Además el papel y la impresión son especiales y costosos.

- el manejo de los formularios debe ser cuidadoso ya que si se dañan no podrán leerse adecuadamente.

- se pierde la posibilidad de la revisión de errores por parte del digitador.

4.5 Detección y Corrección de errores

4.5.1 Descripción General del Proceso

El objetivo de esta etapa es producir un conjunto de datos consistente, que sirva de entrada a los procesos de emisión de tabulación y a la construcción de la base de datos para la diseminación de los resultados.

Para lograr este objetivo se deben evaluar todas las pautas de consistencia con los datos de cada cuestionario determinando un conjunto de inconsistencias que será necesario revisar y corregir.

4.5.2 Tipos de Corrección

En términos generales, la corrección es la asignación de un código válido a una o más de las variables involucradas en la inconsistencia para que la evaluación de la pauta no resulte en error. Entonces, la corrección de la inconsistencia involucra tanto la selección de la(s) variable(s) con errores como también la determinación del código a asignar.

Las correcciones se pueden realizar tanto manual como de forma automática. La corrección manual requiere la revisión de cada uno de los errores por parte de las personas encargadas. La corrección automática la realiza el computador de acuerdo a unas "Reglas de Imputación" que se le suministran.

Es posible efectuar una corrección manual para algunas inconsistencias y posteriormente la corrección automática de las restantes. Por ejemplo, la corrección de errores de

estructura se puede realizar en un proceso manual y las demás inconsistencias de manera automática. Además, opcionalmente la corrección automática puede emitir también un reporte de inconsistencias para las cuales no fue posible su corrección y deberán ser estudiadas manualmente.

La figura 4.4 muestra un diagrama del funcionamiento del proceso de detección y corrección de inconsistencias.

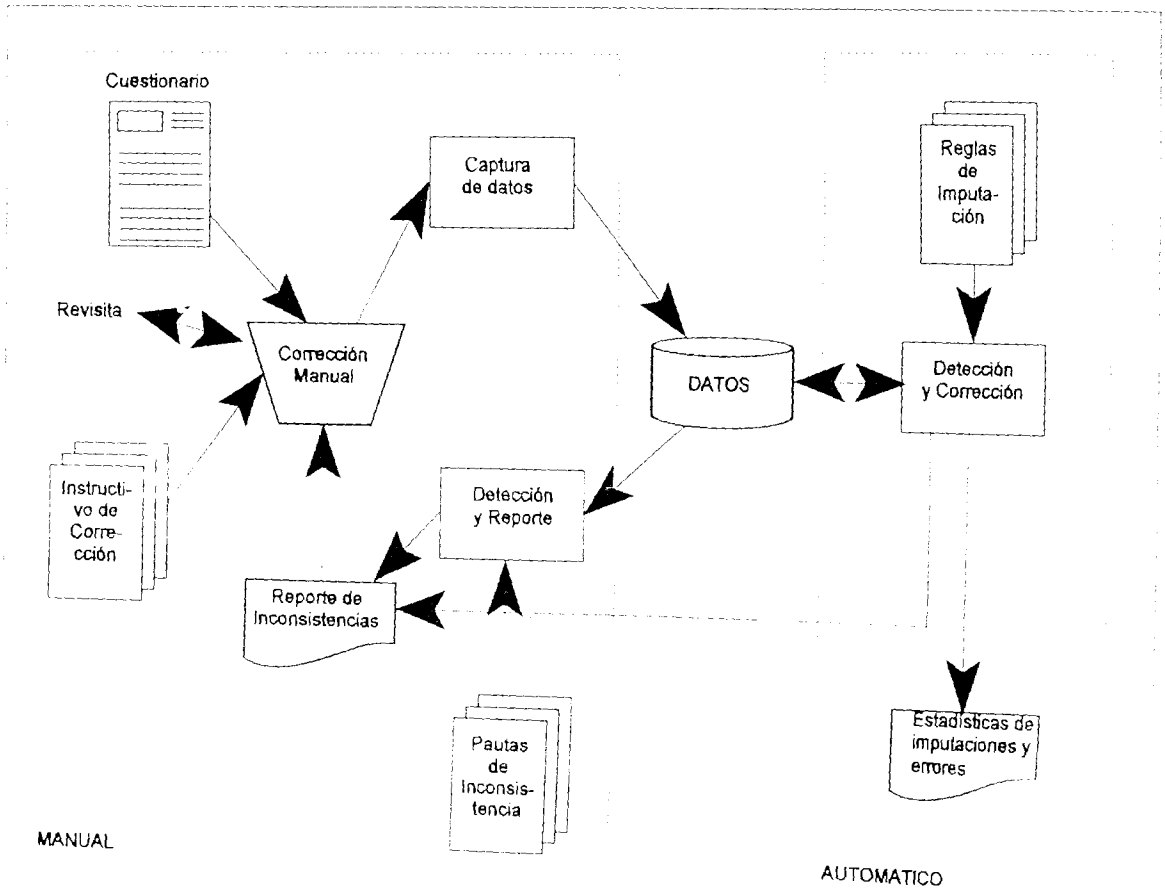


Figura 4.4: Procesos de detección y corrección de errores

Independientemente del mecanismo, la corrección de errores se puede hacer:

- determinando el valor correcto de cada variable por medio del estudio del cuestionario si es un error de digitación o codificación, o regresando al campo (revisita) si se trata de un error en la etapa de recolección.

- asignado un código de "Ignorado" a la variable. Este código debe formar parte del dominio de la variable y debe participar en las pautas de consistencia de tal manera que no produzca mensajes de error.

- estimando un valor de acuerdo a ciertos criterios para que no se de el error pero que tampoco se altere la calidad de la información. Esta técnica es conocida como imputación y es detallada más adelante.

Las decisiones de como realizar las correcciones, que metodología utilizar, etc. deben ser tomadas por los estadísticos, demógrafos o los responsables de la investigación. La labor del informático se debe limitar a apoyar esta tarea de decisión y a preparar el soporte computacional para la ejecución de los procesos.

4.5.3 Corrección Manual

La corrección manual es en la gran mayoría de los casos la alternativa que permite obtener una mejor calidad de los datos pues se hace una revisión detallada de los errores para determinar las causas y hacer las correcciones. Sin embargo, esto no siempre es posible y en muchos casos se debe optar por la corrección automática. Por el contrario, algunas inconsistencias como las de estructura, generalmente requieren de un tratamiento manual ya que no es posible asignar automáticamente toda los valores de los atributos de la ocurrencia de una entidad.

La corrección manual es un proceso recurrente en el que se detecta, reporta, analiza y corrigen errores, y que se ejecuta tantas veces como sea necesarios hasta que no se produzcan más inconsistencias[8].

A continuación se detallan cada uno de los componentes de la corrección manual:

- detección y reporte. Este es un proceso "batch" en el que se evalúan todas las pautas de consistencia con los datos de cada cuestionario, y en el caso de encontrarse algún error, se incluye el cuestionario en el reporte de inconsistencias. El conjunto de datos de entrada a este proceso es el obtenido de la captura, y es recomendable que se utilice en el mismo orden.

---

[8] Es posible que a pesar que existan algunas inconsistencias, como por ejemplo, las pautas *suaves*, se decida que los datos están limpios.

- reporte de inconsistencias. Si un cuestionario presenta al menos un inconsistencias se debe imprimir la identificación del cuestionario, además el número y la descripción de cada pauta de consistencia que resultó errónea. Es necesario también presentar para cada cuestionario inconsistente el valor de las variables involucradas en los errores. Puede ser de provecho imprimir al final del reporte (o cortes parciales) frecuencias con los totales de cada pauta de inconsistencia y otras "estadísticas" similares.

- análisis y corrección. Se debe revisar secuencialmente el reporte de inconsistencias y para cada cuestionario analizar la pauta de consistencia, los datos en el computador y el contenido físico del cuestionario para determinar (si es posible) el error que causa el mensaje de inconsistencia. Si es necesario y factible se puede hacer la "revisita". Una vez determinados los códigos adecuados se debe hacer la corrección tanto en los datos en el computador como en el cuestionario.

- instructivo de corrección. Este documento contiene para cada pauta de consistencia, su número, la descripción y la condición lógica para la cual se dispara. Además es necesario que se incluya una guía que ayude a la persona encargada en resolver cada una de las pautas.

- bitácora de errores/correcciones. Como recomendación opcional, se puede llevar un registro en el computador que contenga los errores emitidos y las correcciones realizadas.

4.5.4 Corrección automática

La corrección automática parte del hecho de que es posible corregir un error determinando un valor adecuado con base en los otros datos del cuestionario o del conjunto de casos sin afectar la "validez estadística" de los resultados.

Entonces, cuando se detecta un error el computador utiliza las "Reglas de Imputación" para seleccionar la variable a corregir y asignarle el código estimado.

Es conveniente usar la corrección automática cuando:

- no es posible realizar la corrección manual, por ejemplo cuando el conjunto de datos es sumamente grande como en los censos de población, o sencillamente cuando no se dispone de los recursos necesarios.

- la proporción de cuestionarios con error es muy baja, de tal forma que la imputación no afecte los resultados ni se produzcan datos "sintéticos".

- no se aceptan inconsistencias y todos los errores deben ser corregidos.

- es factible preparar una regla de imputación adecuada para corregir todos los casos en donde una pauta de inconsistencia pueda detectar un error.

Es indispensable la obtención de cifras de control que presenten al final del proceso los totales de casos para cada tipo de inconsistencia y para cada imputación. De esta forma se podrá confirmar, si el volumen de errores permite la corrección automática y sí las imputaciones se están realizando adecuadamente.

Los componentes de este proceso de corrección automática son:

- reglas de imputación. La definición de estas reglas es una tarea bastante compleja y delicada que debe ser realizada en conjunto por el investigador y el informático. Esta tarea consiste en analizar las paútas de consistencia y preparar una lógica que permita hacer la corrección en cada caso. Para esto, el informático debe, en conjunto con el criterio técnico del investigador, agrupar las pautas de consistencia (probablemente en las secciones de cuestionario), y escribir en "pseudocódigo" los bloques de instrucciones que permitan detectar y corregir todas las posibles inconsistencias de cada grupo.

- detección y corrección. Este proceso lo maneja un programa que se ejecuta "en batch" y que trata cada cuestionario como una unidad, permitiendo que se aplique la lógica definida en las pautas de imputación a cada uno de ellos. Como producto de este proceso se obtiene un conjunto de datos consistente y un reporte de los errores e imputaciones realizadas.

- estadísticas de imputaciones y errores. Este reporte es fundamental, pues determina la calidad del proceso de corrección automática. Aquí se deben presentar las frecuencias de aparición de cada error (para confirmar si es sujeto de corrección automática) y la cantidad y forma de imputar cada variable.

4.5.5 Imputaciones

Las imputaciones se clasifican de acuerdo a la forma en que se selecciona el valor para hacer la corrección.

4.5.5.1 Asignación Única

En este tipo de imputación, llamada también determinística, el valor a asignar se determina de acuerdo a los valores de otras variables del mismo caso o cuestionario, y generalmente se puede expresar en forma de un bloque-if (o varios bloque-if anidados):

```
if condición-lógica
        {otro bloque-if} o
        asignación del nuevo valor
else
        {otro bloque-if} o
        asignación del nuevo valor
end
```

en donde, se hace referencia a variables del mismo cuestionario. Por ejemplo, para corregir un cuestionario con un valor de sexo inconsistente de acuerdo la condición de hijos tenidos:

```
if hijos_tenidos > 0
        if sexo # 2                    (sexo=2: mujer)
                sexo = 2
        end
else
        ...
end
```

4.5.5.2 Asignación Variable

En este tipo de imputación se encuentran las técnicas que obtienen el nuevo valor de acuerdo a procedimientos predeterminados, y se distinguen los principales casos:

- asignación aleatoria: cuando se utiliza un número aleatorio "controlado". Por ejemplo, si se desea imputar el salario para los profesionales en área urbana, es posible generar un número aleatorio entre cierto rango.

- asignación alternada: se puede emplear por ejemplo, para imputar la variable sexo cuando no se conozca su valor, entonces un caso se asigna hombre, el siguiente mujer, luego hombre y así sucesivamente:

```
if sexo # [1-2]
        sexo_imputado = 3 - sexo_imputado
        sexo = sexo_imputado
end
```

nótese que este tipo de corrección puede (y debe en muchos casos) utilizarse conjuntamente con la asignación única para no introducir nuevos errores, por lo que en el ejemplo anterior se debe tener cuidado de no asignar sexo masculino a personas con hijos.

- asignación predeterminada: para asignar valores establecidos con anterioridad al procesos de corrección. Por ejemplo, Se conoce el promedio de edad por sexo y estado conyugal para cada área, entonces se alimenta al computador con arreglos que contienen esta edad, por ejemplo:

```
if edad = NULL
        edad = arreglo_edades(sexo,estado_conyugal)
end
```

de esta forma no se alteran la distribución estadística de la edad en los resultados posteriores y no se hace necesario utilizar una categoría adicional de "ignorado" para la edad de la persona.

A estos métodos se les conoce también con el nombre de paquete frío ("cold deck").

4.5.5.3 Paquete Caliente ("Hot deck")

Esta es una de las técnicas de imputación más utilizadas, y parte del hecho de que es posible asignar a una variable con un valor erróneo el valor de la misma variable de otro cuestionario con las mismas condiciones, o sea, con los mismos valores en otras variables pero consistente.

Por ejemplo, se puede imputar el valor de la variable material de las paredes de la vivienda de acuerdo a los valores de los materiales del techo y piso. Entonces, cuando se encuentre un caso con el material de las paredes inválido, es posible imputarlo con el valor de algún otro cuestionario con los mismos materiales de piso y techo. Esto parece bastante lógico, pues es común que las viviendas se construyan con combinación de grupos materiales para el techo-paredes-piso, principalmente si se utiliza el caso previo con las mismas características.

La implementación de esta técnica se puede hacer mediante el uso de arreglos multidimensionales que se actualizan o "calientan" cuando se procesa un caso consistente, y se utilizan cuando se presenta el caso inconsistente. Siguiendo el ejemplo anterior se puede tener una sección de código de la siguiente manera:

```
if paredes # [1-8]
        paredes = pared(piso, techo)
else
        pared(piso,techo) = paredes
end
```

Aquí, el arreglo "pared" tendrá siempre el último valor válido del material de las paredes para cada combinación de techo y piso, y cuando llegue un caso con dicha variable inconsistente (fuera del rango: 1-8) se corrige con el valor del arreglo. Se deben tomar consideraciones especiales para la inicialización del arreglo, probablemente con las ocurrencias más comunes para cada variable.

Vale la pena indicar que el paquete caliente sólo debe utilizarse cuando exista una homogeneidad en los datos. El caso de los materiales de construcción de las viviendas es el caso típico, se supone de antemano que las viviendas vecinas se parecen. Por el contrario es peligroso asignar por paquete caliente el tamaño de una finca en un censo agropecuario, pues se puede estar asignando a un pequeño lote el tamaño de la hacienda vecina.

4.5.6 Otros métodos de corrección

Las técnicas descritas anteriormente muestran la forma de realizar las correcciones por medio de reglas de imputación, que son diseñadas específicamente para el conjunto de datos de cada investigación. Pero también es posible efectuar imputaciones de manera

general, utilizando algoritmos para la selección de la variable a imputar y del caso "donador" de la información consistente.

Por una parte, las reglas de imputación, definen explícitamente tanto la variable que se desea imputar como también la técnica para realizar la corrección, pero presentan varios inconvenientes como lo laborioso que resulta su definición y el riesgo de que contengan errores lógicos y alteren inadecuadamente los datos de la investigación.

Por lo anterior es que los métodos automáticos, que se comentan resumidamente a continuación, son una alternativa con muchos ventajas, principalmente para la imputación de datos cuantitativos.

Estos métodos de corrección se basan en el algoritmo F-H (Fellegi & Holt) descrito detalladamente en [Bravo 91] y en [Ortúzar 78].

- selección de la variable a imputar: la idea de este primer paso es agrupar todas las pautas de consistencia (llamadas "edits") que involucran determinada variable, entonces para cada cuestionario se evalúan estas pautas para conocer cuales están inconsistentes, seleccionando para corregir la variable que participe proporcionalmente en más pautas inconsistentes.

- localización del caso "donador". Una vez seleccionada la variable a imputar sigue la determinación del valor a asignar a esta variable. Generalmente se puede usar la técnica del "hot deck" descrita anteriormente. La diferencia aquí es la localización del mejor caso de donde tomar el nuevo valor. Para esto se utilizan las variables que participan en las pautas pero que no son inconsistentes y se busca un caso consistente que tenga las mismos valores para estas variables.

4.6 Producción de Resultados Básicos

4.6.1 Consideraciones Generales

El objetivo primordial del procesamiento de los datos de una investigación como las que aquí se estudian, es emitir el conjunto de resultados requeridos y descritos en la etapa de diseño, de tal forma que el investigador disponga de los datos necesarios para que luego de analizarlos, pude confirmar o no las hipótesis planteadas al inicio del proyecto.

Este proceso de emisión de resultados esta determinado en gran medida por el volumen de datos, ya que para conjuntos pequeños[9] se podrá utilizar algún paquete estadístico o similar con el cual pueda interactuar el investigador para obtener los resultados deseados; o sencillamente se trata de hacer consultas a bases de datos relacionales. Por el contrario,

---

[9] En relación con el computador que se utiliza para el proceso

para investigaciones "grandes" es necesario tener en cuenta algunas consideraciones metodológicas que aquí se detallan.

El principal aspecto a tener en cuenta se refiere a la eficiencia con que se obtienen los resultados, y esto está determinado directamente por las lecturas a los datos.

4.6.2 Descripción del proceso

La entrada al proceso de producción de resultados es el conjunto de datos totalmente consistente o datos limpios obtenidos en el proceso de depuración.

La figura 4.5 muestra un esquema con los pasos que debe seguir el proceso de producción de resultados, que está determinado principalmente por la definición de los resultados hecha por el investigador.

4.6.2.1 Tabulación

Esta primera fase del proceso tiene por objetivo realizar la lectura de los datos de entrada y producir un archivo con los contenidos elementales de cada tabla. En otras palabras, esta fase debe producir para cada tabla el contenido de todas las entradas independientes para el nivel de corte de área más bajo.

Por ejemplo, si se solicita un cuadro de salida por edad en grupos quinquenales y sexo para cada región y subregión del país, este fase debe producir el contenido de las celdas para cada grupo de edad y sexo y para cada subregión del país. El cálculo de los totales de la tabla, para sexo y edad, se realiza posteriormente, así también como la agregación de las subregiones en regiones.

Entonces, el archivo de salida de esta fase tendría una ocurrencia por cada subregión, y una estructura como la siguiente:

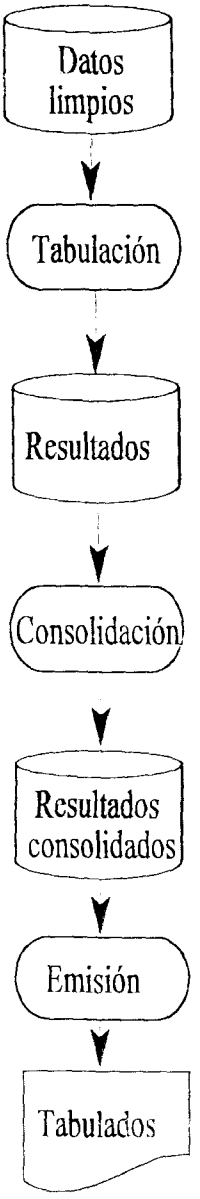


Figura 4.5: Fases en la producción de resultados

```
-identificación de la región
-identificación de la subregión
-contenido de la celda para los hombres del primer grupo de edad
-contenido de la celda para los hombres del segundo grupo de edad
.
.
-contenido de la celda para los hombres del último grupo de edad
-contenido de la celda para las mujeres del primer grupo de edad
.
.
-contenido de la celda para las mujeres del último grupo de edad
```

Nótese que la idea en esta fase es limitar el proceso a las actividades indispensables de lectura y acumulación dejando cualquier otra tarea para las fases posteriores. Esto se debe a que generalmente los archivos de entrada contienen muchos registros, generalmente millones y es necesario minimizar el tiempo que duran estos procesos.

Un concepto importante es el orden del archivo de entrada, y siguiendo con el ejemplo anterior, sería conveniente que el archivo esté ordenado por región y subregión, de modo que cada vez que se cambie de una subregión a otra se grabe el contenido de la tabla.

Para esto es necesario recordar que durante el diseño de los cuadros, el investigador especificó los cortes para los cuales requiere que se emita cada cuadro de salida. El informático deberá estudiar los cortes de área de todos los cuadros, y de esta forma podrá agrupar los cuadros que tengan cortes de área similares. Note que se dice similar y no igual, esto porque un cuadro por región y subregión tiene cortes similares a uno que únicamente se solicite por región.

Si todos los cuadros tiene cortes de área similares sencillamente, se ordena el archivo de entrada en el orden de estos cortes, de manera que se pueda calcular el contenido de cada cuadro fácilmente. Sin embargo esto no siempre es posible, y cuando se tiene cortes distintos se dispone de las siguientes dos posibilidades:

- Realizar tantos procesos como cortes no similares se soliciten. Esto significa que se debe ordenar y procesar el archivo de entrada para cada uno de los cortes de área no similares.

- Realizar un único proceso en el orden de acuerdo a los cortes de área para los cuales se solicite más tabulados. Para los cuadros que requieran otro orden distinto al del archivo de entrada se obtendrán varios resultados en el archivo de salida, que deberán ser consolidados en el fase siguiente.

4.6.2.2 Consolidación

Esta fase del proceso de obtención de resultados tiene la función de consolidar los resultados obtenidos en la fase anterior. Esta consolidación se hace en dos instancias distintas.

Por una parte es necesario sumarizar los resultados obtenidos para los niveles inferiores de cortes de área a cada uno de los niveles superiores. Siguiendo con el ejemplo anterior,

de la fase de tabulación se obtuvo un registro para cada subregión con los datos de sexo y edad, pero la tabla se requiere también para regiones y el total del país. Entonces este proceso suma los resultados de todas las subregiones de cada región y agrega al archivo de resultados estos nuevos registros. Posteriormente, suma los registros de todas las regiones para obtener el registro de resultados a nivel de todo el país.

Como segunda función de esta fase se tiene la consolidación para los cuadros que se tabularon en un orden distinto a lo requerido. Entonces, en esta fase se debe ordenar el archivo de resultados en orden de tabla para tener los registros de una misma tabla juntos y así poder realizar la consolidación correspondiente.

El resultado de este proceso de consolidación es un archivo con un registro para cada tabla y para cada uno de los cortes de área que se solicitaron.

4.6.2.3 Emisión

Una vez que se tiene consolidado el archivo de resultados se puede pasar a la emisión de los cuadros de salida. Esta fase de emisión se compone de dos partes:

- Primero es necesario hacer el cálculo de las entradas dependientes de las tablas. En el mismo ejemplo, se puede calcular la columna correspondiente al total de cada grupo de edad para ambos sexos, y la fila con el total de todos los grupos

de edad para cada sexo. Aquí es necesario considerar todas las anotaciones que haga el investigador en relación al cálculo de las entradas dependientes.

- Posteriormente se puede pasar a la propia emisión de la tabla, en un proceso en el cual se agrega el formato deseado y todos los componentes textuales de la tabla, como el título, los encabezados, el pie de página, etc.

5. ORGANIZACIÓN DE DATOS

5.1 Comentario General

El procesamiento de los datos de un censo o encuesta comprende tanto las actividades de captura, depuración y producción de los resultados pre-definidos y descritas en los capítulos anteriores, como también las referentes al diseño de la o las bases de datos para la diseminación de la información final.

Esta consideración puede llevar a pensar que la organización de los datos debe satisfacer los requerimientos de cada una de la fases del procesamiento. Sin embargo, en la gran mayoría de los casos, esto no es conveniente ni recomendable debido a las grandes diferencias que hay entre cada una de las fases.

De lo anterior, se distingue la existencias tres instancias que son: la captura y depuración de datos, la producción de resultados básicos y la diseminación de los datos finales; cada una con sus propios requerimientos que son distintos entre sí. Por lo tanto, una misma organización de datos no necesariamente debe satisfacer a las tres instancias.

Además de lo anterior, es interesante señalar que únicamente la organización de datos en su tercera instancia permanece ya que la captura y depuración y las tabulaciones básicas tienen una existencia temporal.

5.2 Requerimientos de la organización de datos

A continuación se comenta los requerimientos que presenta cada una de ellas:

5.2.1 Captura y depuración

Como se ha comentado en el capítulo anterior, durante la captura y depuración de datos se requiere una organización de con las siguientes características:

- Debe permitir identificar el caso individualmente, ya sea el propio cuestionario o alguno de sus componentes, o sea las ocurrencias de algunas de las entidades

- en muchos casos, se deben agrupar los datos por área de enumeración, entonces es necesario identificar a que área pertenece cada cuestionario

- Los datos pueden ser procesados de manera descentralizada

- Se requieren las funciones básicas de cualquier sistema como son: la inserción, la modificación e incluso la exclusión, además de la consulta de casos individuales.

Es seguro que los requerimientos anteriores se puedan satisfacer de manera adecuada con una base de datos relacional como las que se establecen comúnmente en sistemas de índole administrativo. Por lo tanto no se estima necesario incluir en esta guía consideraciones metodológicas sobre esta instancia en la organización de los datos.

5.2.2 Tabulaciones básicas

A diferencia de la instancia anterior, la emisión de las tabulaciones básicas presenta características propias. Estas características se detallan a continuación:

- Es un proceso exclusivamente de lectura de los datos y que debe hacerse generalmente en uno o varios ordenes determinados[10]

- Se procesa el conjunto total de los datos a pesar de que es posible realizar el proceso descentralizadamente

- No interesa identificar el caso individualmente, únicamente la relación entre las ocurrencias de las entidades de un cuestionario

---

[10] Definido de acuerdo a la definición de los cortes en el diseño de los cuadros de salida.

Debido a lo anterior es que tradicionalmente se ha realizado este proceso de manera secuencial y sin mayor estructura en los datos que el propio ordenamiento de acuerdo a los cortes descritos para los cuadros de salida.

Esta consideración es perfectamente válida pues evita el "overhead" en el acceso a los datos producto de las estructuras internas en las bases de datos.

Sin embargo, la decisión de como organizar los datos para la emisión de los resultados básicos depende al final de cuentas de otras consideraciones, como:

- el volumen de casos

- la cantidad de distintos cortes de área similares

5.2.3 Diseminación

Es posible considerar la diseminación de datos como una última etapa en el procesamiento de los datos de censos y encuestas, y que en muchos casos ha sido muy poco aprovechada o incluso totalmente ignorada, lo que ha llevado a una gran subutilización de la información proveniente de estas investigaciones. Lo anterior es

probable que tenga como causa la propia complejidad o la gran cantidad de recursos que son requeridos.

La diseminación de los datos pretende poner a disposición de los usuarios la totalidad o parte de la información producida por la investigación, en una o varias organizaciones distintas, de tal forma estos puedan satisfacer la mayoría de sus requerimientos. Es probable también encontrar aplicaciones en que se requiera datos provenientes de distintas fuentes, como por ejemplo, censos de distintos años.

Es posible preparar aplicaciones para la diseminación de los datos en dos formas, elemental o agregada. El dato elemental o también llamado "micro-dato" se refiere al conjunto de datos el mismo nivel de su captura, por el contrario, por dato agregado o "macro-dato" se entiende la presentación del dato elemental sumarizado a otros niveles distintos al original.

Por ejemplo, en un censo de población y vivienda, el dato elemental se refiere a la información de cada una de las viviendas y de las personas que la forman, por el contrario, el dato agregado podría ser el número de personas por sexo y grupo de edad para determinado ubicación geográfica.

Ambas presentaciones de datos tiene sus propias características y utilidades, el dato agregado es de más fácil uso de parte del usuario pero con limitaciones en el análisis,

y con el dato elemental es posible utilizar la totalidad de la información pero con una mayor complejidad. A continuación se comentan ambas representaciones.

5.2.3.1 Datos agregados

De acuerdo al tipo de investigación que origine los datos, es probable encontrar bases de datos agregadas de distintos tipos. Sin embargo, existen agregaciones de datos que se utilizan con mayor frecuencia, como por ejemplo:

Tablas estadísticas

Estas son bases de datos que contienen una gran cantidad de tablas previamente elaboradas y se espera que los requerimientos del usuario sea satisfechos con estas tablas. La gran ventaja de estos sistemas es que el tiempo de recuperación de una tabla es insignificante debido a que no tiene que realizar cálculo alguno.

El sistema debe organizar las tablas de modo que facilite al usuario su recuperación. Por ejemplo, las tablas pueden agruparse por área geográfica y nivel de desagregación, además debe tener vías de acceso de acuerdo a las variables contenidas en la tabla.

Sistemas de Información Geográficos

El desarrollo de los Sistemas de Información Geográficos (SIG) ha abierto un gran campo para la diseminación de la información de censos y grandes encuestas. Ahora, las bases de datos con información estadística deben proveer los datos agregados a los SIG's para que el usuario pueda preparar sus mapas temáticos para las áreas geográficas y niveles deseados.

Como caso particularmente importante se encuentran los Censos de Población y Vivienda que ofrecen la información poblacional que puede servir de base a una gran mayoría de aplicaciones de los Sistemas de Información Geográficos.

Sumarización de datos

Otro campo importante en las bases de datos con información agregada, es la simple sumarización de variables a distintos niveles jerárquicos, dentro de la estructura política administrativa o institucional de los países. Por ejemplo, se pueden preparar bases de datos con la información demográfica básica agregada para cada uno de las áreas de enumeración o segmentos censales.

5.2.3.2 Datos elementales

Este tipo de bases de datos para la diseminación de la información de censos y encuestas presenta requerimientos particulares que la diferencias de la mayoría de las bases de datos que se utilizan comúnmente. A continuación se comentan estos requerimientos:

- No debe ser posible identificar el caso individualmente (ver apartado sobre el secreto estadístico más adelante)

- Se requieren únicamente procesos de recuperación de datos y nunca de actualización, ya que los datos almacenados son estáticos

- En la gran mayoría de ocasiones se procesan sólo subconjuntos del total. Es muy probable que la gran mayoría de las consultas se realicen sobre selecciones del conjunto de casos y en muy pocas ocasiones sobre el total

- Las consultas se realizan para pocos atributos de muchas ocurrencias.

De lo anterior se llega a que las bases de datos para estos fines deben tener algunas características particulares como las que se detallan a continuación:

Organización jerárquica

El muchas partes de esta guía se ha mencionado la estructura jerárquica que presentan las entidades en el cuestionario. Además, ahora hay que agregar que el acceso a estos datos también se da sobre una organización jerárquica.

La selección de los subconjuntos de datos para la obtención de resultados se hace en la gran mayoría de las ocasiones en la estructura geográfica o institucional de los países o regiones en los cuales se realizó la investigación.

Por ejemplo, una consulta puede darse para conocer las características demográficas del territorio cercano a un río contaminado. Entonces, el sistema debe permitir al usuario seleccionar las áreas geográficas adyacentes al río y emitir para este subconjunto de casos los resultados pedidos.

Unido a lo anterior, los países o instituciones se organizan jerárquicamente en donde cada división está a su vez subdividida en otro nivel jerárquico.

Las implicaciones de esta característica se dan principalmente en la organización física de los datos, ya que es posible cargar la base de datos en un orden físico similar a la estructura de acceso, estableciendo punteros que determinen la posición relativa del área deseado en el archivo total de datos.

Archivos invertidos

El procesamiento estadístico de datos, a diferencia de otros temas de la informática, requiere generalmente una pocas variables de una gran cantidad de casos para un proceso dado.

Por ejemplo, si se desea conocer la distribución de la población por edad y sexo, es necesario leer únicamente las variables EDAD y SEXO y no los demás atributos de la entidad POBLACION. Con esto se logra minimizar los accesos a disco y solamente traer a memoria las variables deseadas.

Para tal efecto, se considera como una buena alternativa para la organización física de los datos el uso de archivos invertidos en lugar de los esquemas tradicionales de almacenamiento.

De esta forma, se tendrían almacenadas conjuntamente todas las ocurrencias de un atributo en vez de todos los atributos de una ocurrencia. Entonces, a la hora de realizar una consulta, cada lectura de la base de datos obtiene un cantidad determinada de ocurrencias de la variable deseada.

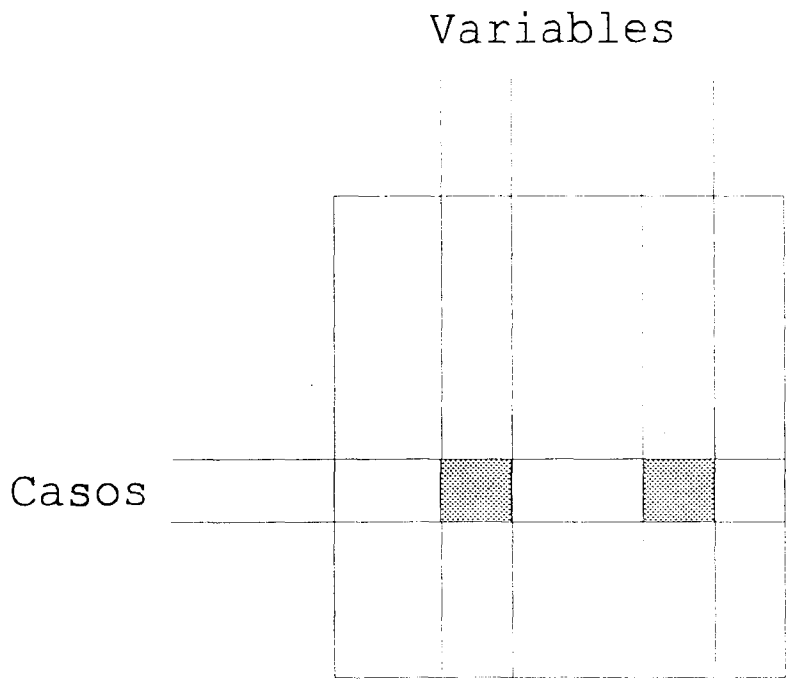


Figura 5.1: Acceso a únicamente los datos requeridos

Como comentario se puede asegurar que una estructura interna con las características mencionadas permite que los procesos recuperen de disco únicamente los valores para las variables deseadas y para los casos seleccionados, por lo que no se incurre en desperdicio alguno de recursos llevando a memoria información que no se utilizará. En la figura 5.1 se muestran las áreas sombreadas, las cuales son las que se traerán del disco.

El secreto estadístico

Finalmente es necesario hacer un comentario importante referente al secreto estadístico.

En la gran mayoría de los países existe legislación que protege al informante sobre el destino de la información que suministra. Esto motiva a que el informante colabore con la recolección de los datos dando información completa y verdadera.

Esta legislación conocida como "secreto estadístico" obliga a las autoridades encargadas de las estadísticas a evitar que terceras personas conozcan los datos suministrados por el informante. Esto se convierte en una gran limitante a la hora de diseñar las bases de datos, y deben tomarse las siguientes consideraciones:

- no debe existir una identificación para cada caso en la base de datos. El otras palabras, no debe ser posible acceder a un caso en particular

- los accesos a la base de datos no deben permitir que el usuario obtenga niveles tan desagregados de datos que le permita hacer aproximaciones a la entidad de las personas

CONCLUSIONES

La mejor manera de lograr que el procesamiento de los datos de un censo o encuesta tenga una buena probabilidad de éxito, es siguiendo consideraciones metodológicas adecuadas a las características de ese tipo de procesamiento. La contribución de este trabajo consiste precisamente en el desarrollo de una guía metodológica que tome en cuenta dichas características.

El ignorar las particularidades del procesamiento de censos y tratarlo como si tuviera las características comunes a otro tipo de sistemas, atenta contra la calidad de los resultados y la eficiencia al obtenerlos. No es extraño escuchar comentarios como que cierto análisis no se pudo realizar porque las variables se recogieron de manera inadecauda o que un cuestionario mal diseñado provocó una tasa de errores o valores faltantes mucho más alta de lo permitido.

La realización de este trabajo produjo como resultado una metodología concreta y detallada con las consideraciones para cada etapa del procesamiento, desde la propia formulación de la investigación hasta la producción de los resultados. Además, se obtuvo también la información relevante tanto para que el informático como el investigador se enfrenten adecuadamente al procesamiento de datos de censos y encuestas.

Es necesario considerar que estos aspectos metodológicos no son del todo permanentes, ya que están fuertemente condicionados por los avances tecnológicos. Por ejemplo, nuevos dispositivos para la recolección de datos o una reducción significativa en los costos de ciertos equipos, deben motivar a los usuarios de esta guía a revisar las recomendaciones aquí descritas tratando de adecuarlas a estas mejoras. Pero también es claro que los países subdesarrollados deberán asimilar el desarrollo tecnológico de acuerdo a sus propias posibilidades, aceptando principalmente las mejoras que se adapten a una buena relación de costo/beneficio.

En otro orden de cosas, esta guía no resulta exhaustiva, ya que existen varios temas para los cuales se podría profundizar su estudio. Dentro de estos temas, que requeriría cada uno de un documento independiente se encuentra:

- Aspectos computacionales en la corrección automática de errores. Este tema es lo suficientemente extenso y complejo que permite al informático colaborar con la resolución de este tipo de problemas que ha sido tratado principalmente por profesionales en el área de la estadística y similares.

- Consideraciones metodológicas para la diseminación de datos de encuestas y censos. Aquí se trataría principalmente en desarrollar una metodología para el paso de los datos depurados a los medios de diseminación como las bases de datos estadísticas, los Sistemas de Información Geográficos, sitios en internet, etc.

Para terminar, resulta particularmente interesante el desarrollo de un sistema CASE o de ingeniería de software asistido por computador, que ayude tanto al investigador como al informático en el procesamiento de los censos y encuestas.

El apoyo de este CASE comienza con las herramientas necesarias para que el investigador diseñe las tablas con los resultados requeridos. Posteriormente, el sistema facilitaría la labor del modelaje de datos al informático obteniendo como resultado un diccionario de datos.

También podría colaborar con el diseño del cuestionario preparando un esbozo gráfico del formulario incluyendo las variables con las características propias a su tipo, las secciones, los saltos, etc.

Finalmente, y debido a que los programas de entrada de datos, detección y corrección de errores y emisión de tabulaciones tienen una lógica bastante predefinida, el CASE podrá fácilmente preparar versiones casi definitivas de estos programas.

APÉNDICES

1. Un Ejemplo de la metodología[11]

Censo Nacional de Escolares

Evaluación del Rendimiento en Matemáticas

Descripción del Problema

El rendimiento en matemáticas que han obtenido los alumnos de primaria durante el año anterior ha sido deficiente, confirmando la tendencia decreciente de los últimos períodos. Esta situación ha causado mucha preocupación dentro de las autoridades educativas del país, por lo que se ha decidido realizar una investigación con el objetivo de determinar las causas de dicho problema y buscar algunas alternativas de solución.

Las posibles causas de este bajo rendimiento, según los expertos, abarcan una serie de aspectos para los cuales se necesita conocer sus características y su situación actual. Dentro de estos aspectos se menciona el nivel socioeconómico y las cualidades personales de los alumnos, la capacitación del profesorado, las características físicas del aula, y otros más.

---

[11] Este ejemplo ha sido preparado con el principal objetivo de mostrar la utilidad de esta metodología por lo que desde el punto de vista de la investigación sociológica y educativa su contenido puede no ser el más adecuado.

Debido a que no existe la información necesaria para hacer los análisis que confirmen o rechacen las hipótesis planteadas, se ha decidido realizar un "Censos Nacional de Escolares" que sirva para recolectar los datos y posteriormente producir los resultados requeridos.

Luego de definir un marco teórico, y plantearse las hipótesis del caso, los investigadores a cargo del censo diseñaron el siguiente conjunto de cuadros, que esperan obtener de la investigación y que servirán para comprobar las hipótesis:

Definición de los resultados requeridos (apartado 3.2.2)

Cuadro 1. Cantidad de alumnos y promedio de matemáticas por nivel socio-económico según Región, circuito, subcircuito, sexo y grupos de edad

Región, circuito, subcircuito, sexo y grupos de edad	Total		Alto		Medio		Bajo	
	Alumnos	Promedio	Alumnos	Promedio	Alumnos	Promedio	Alumnos	Promedio
	(1) = (3)+(5)+(7)	(2) = [(4)+(6)+(8)]/(1)	(3)	(4) = (4)/(3)	(5)	(6) = (6)/(5)	(7)	(8) = (8)/(7)

Ambos Sexos (1) = (2)+...+(5)
- 5 o menos (2) = (7)+(12)
- 6 - 8 (3) = (8)+(13)
- 9 - 11 (4) = (9)+(14)
- 12 y más (5) = (10)+(15)

Hombres (6) = (7)+...+(10)
- 5 o menos (7)
- 6 - 8 (8)
- 9 - 11 (9)
- 12 y más (10)

Mujeres (11) = (12)+...+(15)
- 5 o menos (12)
- 6 - 8 (13)
- 9 - 11 (14)
- 12 y más (15)

U = {todos los casos}

Cortes

0	Total	S
1	Región	S
2	Circuito	S
3	Subcircuito	-

Nota: Para calcular este cuadro se acumula en las columnas de "Promedio" la suma de las calificaciones de matemática de cada alumno, luego se hacen los totales tanto de las columnas como de las filas y posteriormente se dividen entre el número de alumnos para obtener los promedios.

Cuadro 2. Cantidad de alumnos y promedio de matemáticas por condición de repitencia según Región, nivel académico y experiencia del maestro

Región, nivel académico y experiencia del maestro	Total		Repitentes		No Repitentes	
	Alumnos	Promedio	Alumnos	Promedio	Alumnos	Promedio
	(1) = (3)+(5)	(2) = [(4)+(6)]/(1)	(3)	(4) = (4)/(3)	(5)	(6) = (6)/(5)

Todos (1) = (2)+...+(5)
- ninguna (2) = (7)+(12)+(17)
- 1-2 años (3) = (8)+(13)+(18)
- 3-4 años (4) = (9)+(14)+(19)
- 5 anos y más (5) = (10)+(15)+(20)

Sin estudios (6) = (7)+...+(10)
- ninguna (7)
- 1-2 años (8)
- 3-4 años (9)
- 5 años y más (10)

Incompletos (11) = (12)+...+(15)
- ninguna (12)
- 1-2 años (13)
- 3-4 años (14)
- 5 años y más (15)

Completos (16) = (17)+...+(20)
- ninguna (17)
- 1-2 años (18)
- 3-4 años (19)
- 5 años y más (20)

U = {todos los casos}

Cortes

0	Total	S
1	Región	-

Nota: véase nota del cuadro 1.

Cuadro 3. Cantidad de alumnos y promedio de matemáticas por nivel de hacinamiento según Región, circuito, sexo y grupos de edad

Región, circuito, y subcircuito, sexo y grupos de edad	Total		Alto		Medio		Bajo	
	Alumnos	Promedio	Alumnos	Promedio	Alumnos	Promedio	Alumnos	Promedio
	(1) = (3)+(5)+(7)	(2) = [(4)+(6)+(8)]/(1)	(3)	(4) = (4)/(3)	(5)	(6)= (6)/(5)	(7)	(8)= (8)/(7)

Ambos Sexos (1) = (2)+...+(5)
- 5 o menos (2) = (7)+(12)
- 6 - 8 (3) = (8)+(13)
- 9 - 11 (4) = (9)+(14)
- 12 y más (5) = (10)+(15)

Hombres (6) = (7)+...+(10)
- 5 o menos (7)
- 6 - 8 (8)
- 9 - 11 (9)
- 12 y más (10)

Mujeres (11) = (12)+...+(15)
- 5 o menos (12)
- 6 - 8 (13)
- 9 - 11 (14)
- 12 y más (15)

U = {todos los casos}

Cortes

0	Total	S
1	Región	S
2	Circuito	S
3	Subcircuito	-

Nota: Refiérase a la nota el cuadro 1. El hacinamiento es una variable derivada y debe ser calculada de la siguiente forma: se divide el área total del aula en metros cuadrados entre el número de alumnos de la sección; si el resultado es menor de 0.33 alumnos por m2 el hacinamiento es bajo, si está entre 0.33 y 0.5 es medio, y si es mayor de 0.5 es alto.

Cuadro 4. Cantidad de alumnos con baja calificación y promedio de matemática por condición de repitencia según Región, ubicación de la escuela y tipo

Región, ubicación y tipo de escuela	Total		Repitentes		No Repitentes	
	Alumnos	Promedio	Alumnos	Promedio	Alumnos	Promedio
	(1) = (3)+(5)	(2) = [(4)+(6)]/(1)	(3)	(4) = (4)/(3)	(5)	(6)= (6)/(5)

Total (1)=(2)+...+(5)
- Oficial (2)=(6)+(10)
- Semi-Oficial (3)=(6)+(11)
- Privado (4)=(7)+(12)

Urbano (5)=(6)+...+(8)
- Oficial (6)
- Semi-Oficial (7)
- Privado (8)

Rural (9)=(10)+...+(12)
- Oficial (10)
- Semi-Oficial (11)
- Privado (12)

U = {alumnos con calificación < 60}

Cortes

0	Total	S
1	Región	-

Nota: véase nota del cuadro 1.

Los investigadores solicitan además que el cuadro 3 se emita cambiando la variable de hacinamiento por iluminación del aula y luego por ventilación. Ambas variables pueden tener como categorías: mala, regular y adecuada.

Además, se considera conveniente contar con una variable a nivel de escuela que indique el tipo de horario en que se imparten las lecciones: matutino, vespertino o mixto.

Para las escuelas rurales se ha solicitado determinar si el acceso de los estudiantes al centro educativo es fácil o difícil, además de la característica de la zona como rural concetrada o rural dispersa.

Características de la investigación

El Ministerio de Educación ha dividido el país en Regiones de Educación. Cada Región a su vez esta subdividida en Circuitos Escolares y estos en Subcircuitos. Cada centro educativo o escuela pertenece a un subcircuito determinado. La identificación de cada nivel de esta estructura se hace dentro del nivel superior, por ejemplo, los circuitos se numeran dentro de cada región a partir de uno. De la misma forma, las escuelas tienen un código numérico asignado dentro de cada subcircuito.

Por su parte, las escuelas desarrollan sus programas en niveles que van del primero al sexto. Es común que en cada grado se tengan varias secciones o grupos no más de cinco

por nivel, lo que lleva a una numeración como 3-2 para indicar la sección 2 del tercer grado. Los alumnos se numeran de uno en adelante dentro de la sección y en orden alfabético.

Por su parte, el ciclo educativo está dividido en tres trimestres cuya nota se promedia para obtener la nota final del curso. La calificación del trimestre se obtiene de un examen trimestral, exámenes parciales, tareas y la evaluación subjetiva de la maestra.

La recolección de datos se realizará durante el curso lectivo, por lo que el rendimiento de cada alumno se medirá con su calificación del primer trimestre.

Identificación de las variables y entidades (apartado 3.3)

A continuación se presenta la lista de variables y las características que el informático obtuvo de los cuadros que diseño el investigador:

Variable	Tipo	Dominio	Entidad
Región	N	1-6	escuela
Circuito	N	1-5	escuela
Subcircuito	N	1-10	escuela
horario	N	1-matutino 2-vespertino 3-mixto	escuela
Calificación de matemáticas	C	0-100	alumno
nivel socio-económico	N	1-Alto 2-Medio 3-Bajo	alumno
Sexo	N	Masculino Femenino	alumno
Edad	C	5-15 años	alumno
Repitencia	N	1-Sí 2-No	alumno
Nivel académico del maestro	N	0-sin estudios 1-incompletos 2-completos	sección
Experiencia docente del maestro	N	0-ninguna 1 a 5 años	sección
área del aula	C	1-100 m2	sección
iluminación	N	1-mala 2-regular 3-adecuada	sección
ventilación	N	1-mala 2-regular 3-adecuada	sección
zona de la escuela	N	1-urbana 2-rural	escuela

tipo de escuela	N	1-oficial 2-semioficial 3-privada	escuela
código de escuela	N	1-20	escuela
grupo de la sección	N	1-5	sección
número del alumno	N	1-40	alumno
nivel	N	1-6	sección
acceso	N	0-NA 1-fácil 2-difícil	escuela rural
ruralidad	N	0-NA 1-concetrada 2-dispersa	escuela rural

(N-numérica categorizada, C-numérica contínua)

Comentarios:

- se informó al informático que existen listados de las regiones, circuitos y subcircuitos.

- para calcular el hacinamiento ha sido necesario incluir la variable "dimensión" del aula.

- inicialmente se consideraron entidades para el maestro, el aula y la sección, pero como se descubrió que cada aula tiene una única sección y este un único maestro, se decidió considerar solamente la entidad sección.

- el informático asignó códigos numéricos a las categorías de las variables que lo requerían, como sexo, condición de repitencia, etc.

- se han agregado las últimas cuatro variables para completar la identificación de las entidades. Para la entidad "sección" se utilizan dos variables, el nivel y la sección o grupo dentro de ese nivel.

- las variables acceso y ruralidad tienen códigos de No aplica (NA) ya que pertenecen a la subclase de escuelas rurales.

Diagrama de entidad-relación (apartado 3.3)

El informático preparó el diagrama de entidad-relación mostrado en la figura 1.1 para facilitar la comprensión del modelo de datos.

Las entidades "sección" y "alumno" son débiles pues únicamente se identifican totalmente dentro de "escuela" y "sección" respectivamente.

Nótese la subclase de escuelas rurales con sus propios atributos.

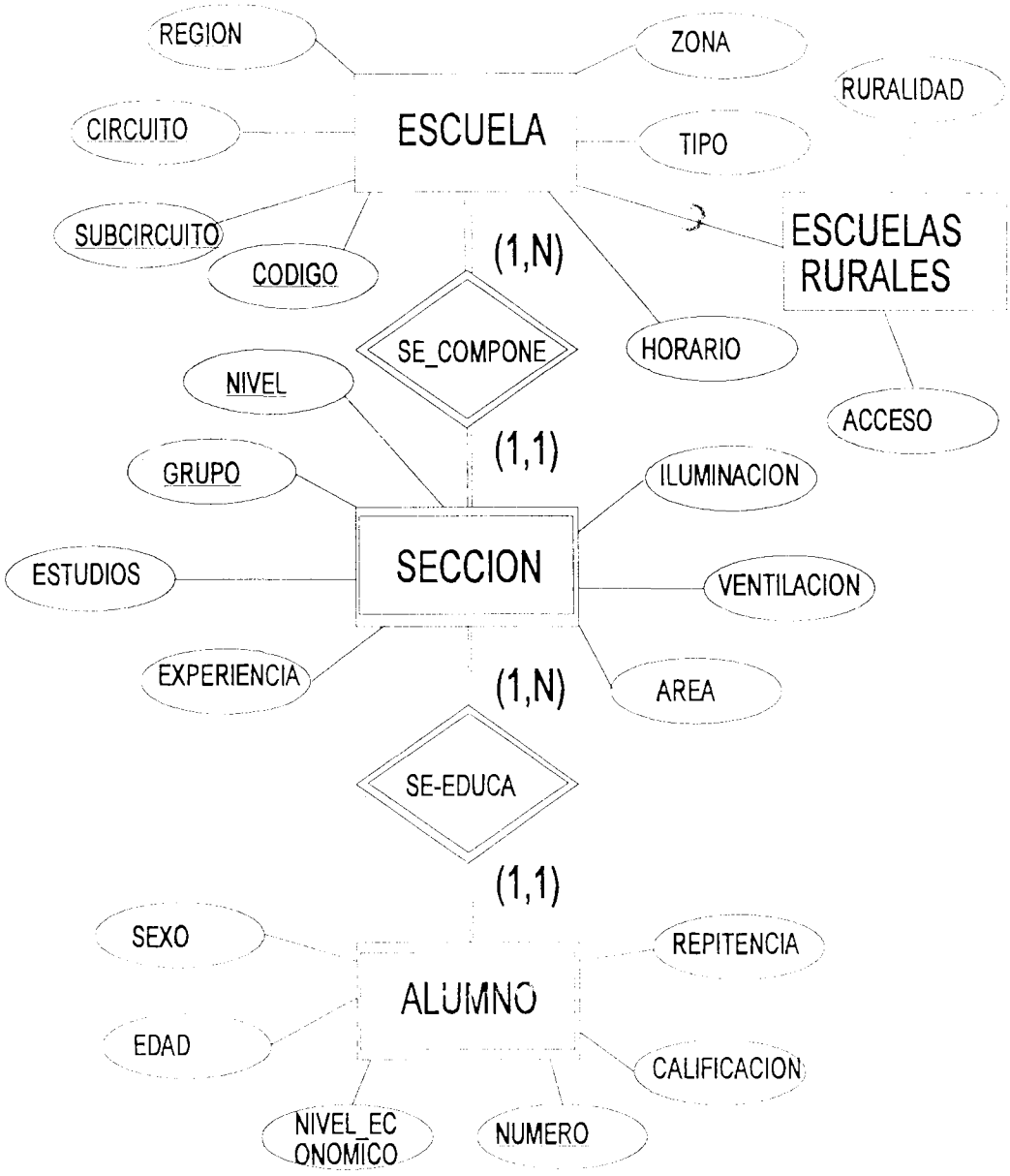


Figura 1.1: Diagrama de entidad-relación del ejemplo

Pautas de consistencia (apartado 3.3.3)

De acuerdo a las características de los datos y a la opinión del investigador principal, el informático preparó la siguiente lista de pautas de consistencia. Cada pauta muestra una condición de error, y se muestran primero las pautas para la revisión del dominio de las variables:

region # [1-6]

circuito # [1-5]

subcircuito # [1-10]

escuela # [1-20]

tipo # [1-3]

zona # [1-2]

grupo # [1-5]

nivel # [1-6]

estudios # [0-2]

experiencia # [0-5]

area # [1-100]

numero # [1-40]

sexo # [1-2]

edad # [5-15]

repitencia # [1-2]

calificacion # [0-100]

nivel_economico # [1-3]

accesso # [0-2]

ruralidad # [0-2]

Algunas pautas interesantes que permite hacer algunas revisiones de la consistencia de los datos:

-la edad mínima para cada nivel es 5 años más que la edad del niño:

edad < nivel + 5

-la cantidad de metros cuadrado para cada niño es demasiado pequeño (menor o igual a 1 metro):

area / TOTAL(alumno) <= 1

-si se supone que las escuelas privadas no contratan maestras sin experiencia ni estudios:

tipo = 3 y experiencia = 0 y estudios = 0

-las variables que miden el acceso las escuelas y la ruralidad se hacen únicamente a las escuelas rurales:

(zona = 2 Y (acceso = NA O ruralidad = NA)) O
(zona = 1 Y (acceso # NA O ruralidad # NA))

Cuestionario y diseño del formulario (apartado 3.4)

En este caso se dispone de dos formularios. Uno para la escuela que lo deberá completar su director y otro para cada una de la secciones que lo llenará el maestro del grupo. Posteriormente a la recolección, el encargado del censo en la escuela procederá a juntar todos los formularios encabezados por el propio de la escuela. El conjunto de todos estos formularios conforma el cuestionario (lógico).

De acuerdo a la definición de los datos, el cuestionario está identificado por la composición de los códigos de la Región, el Circuito, el Subcircuito y el de la propia escuela. A esto se le agrega los códigos del nivel y grupo para identificar a la sección, y el número de alumno que identifica a cada uno de los estudiantes.

En el cuestionario se observa claramente la existencia de tres grandes partes relacionadas con las tres entidades presentes en la investigación: "escuela", "sección" y "alumno".

A continuación se presenta el diseño del cuestionario

MINISTERIO DE EDUCACION PUBLICA

CENSO DE ESCOLARES

RENDIMIENTO EN MATEMATICAS

REGION _____ ☐

CIRCUITO _____ ☐☐

SUBCIRCUITO_____ ☐☐

ESCUELA _____ ☐☐☐

Fecha: ☐☐☐☐☐☐

día mes año

DATOS ESCUELA

Zona:

O Urbana 1

O Rural 2

Tipo:

O Oficial 1

O Semi-oficial 2

O Privada 3

Horario:

O Matutino 1

O Vespertino 2

O Mixto

Escuelas rurales

Accesibilidad:

O Fácil 1

O Difícil 2

Ruralidad:

O Concentrada 1

O Dispersa 2

Observaciones:

_____

_____

_____

_____

_____

_____

_____

Director: _____ Encargado censo: _____

Escuela: _____

Nivel: _____ ☐ Grupo: _____ ☐

DATOS SECCION

Maestro

Estudios: O 1 Sin estudios
O 2 Incompletos
O 3 Completos

Experiencia: ☐☐ años

Aula

Area total: ☐☐ m2

Iluminación:
O 1 Mala
O 2 Regular
O 3 Adecuada

Ventilación:
O 1 Mala
O 2 Regular
O 3 Adecuada

DATOS ALUMNOS

#	Nombre	Edad	Sexo	Repite	Nivel Econom.	Nota Matem.
		años	1-varón 2-mujer	1-Sí 2-No	1-Alto 2-Medio 3-Bajo	
1						
2						
3						
4						
5						
6						
7						
8						
9						
:						
:						

Captura y procesamiento de datos (capítulo 4)

Se utilizará la estructura administrativa del Ministerio de Educación para la recolección de datos. Por tal motivo, en cada subcircuito, circuito, región y a nivel central se dispondrá de personal encargado del trasiego y control de los formularios.

Como la información será recolectado por personal calificado como los directores y maestros se supone una altísima calidad en los datos y no se establecerán controles ni supervisión en el campo.

A nivel central se dispondrá de un cuerpo de revisores o críticos que darán una revisada general a cada cuestionario para determinar errores que afecten la captura de datos, como la identificación y las respuestas múltiples. Debido a que los formularios tendrán impresos los códigos de identificación con anterioridad no existen tareas de codificación.

La captura de datos se hará digitando la información en microcomputadores provistos de un programa elemental que chequea únicamente rangos de variables. No habrá verificación.

Debido a que se dispone de muy fácil acceso a los centros educativos vía telefónica, todos los errores detectados serán impresos para corregirlos manualmente previa consulta al campo.

Para efectuar las tabulaciones básicas, el archivo de datos limpios será clasificado en el orden de región, circuito, subcircuito ya que todas los tabulados tienen esos cortes similares.

2. Programas para el procesamiento de censos y encuestas

2.1 IMPS: Integrated Microcomputer Processing System

2.1.1 Introducción

El IMPS es un programa desarrollado y mantenido por el International Programs Center (IPC) of the U.S. Bureau of the Census, y es utilizado por las oficinas de censos de muchos países, principalmente del tercer mundo.

A diferencia de otros programas, el IMPS no ha sido desarrollado como una única pieza de software, sino por el contrario, ha sido el producto de la unión de algunos programas independientes que se habían venido usando desde principios de la década de los setentas. Estos programas, el CENTS para la emisión de tabulaciones, y el CONCOR para la detección y corrección de errores fueron las principales herramientas disponibles para el procesamiento de censos y encuestas en los "mainframe" de la época.

Durante la década de los ochentas, el IPC adapta estos programas para que funcionaran en microcomputadores PC-compatibles. Posteriormente, aparece el IMPS en donde se integra el CONCOR y el CENTS con un diccionario de datos, un programa para la entrada de datos y otros programas menores, que conformaron un paquete que permite satisfacer las necesidades básicas de los usuarios en esta área del procesamiento de datos.

2.1.2 Características básicas

La figura 2.1 muestra los principales componentes del IMPS y las relaciones entre ellos, así también como su participación en el procesamiento de los datos de un censo o encuesta.

El punto central del sistema consiste en un diccionario de datos que es administrado por uno de sus módulos llamado **DATADICT**. Este diccionario contiene la definición de los datos que se van a procesar y sirve de base para cada uno de los otros módulos.

Para la entrada de datos se dispone de un programa llamado **CENTRY** que permite la construcción de una aplicación de captura de datos y su posterior ejecución. Específicamente, este programa puede ser usado para capturar, revisar, verificar y modificar los datos y además sirve para obtener estadísticas sobre el rendimiento del digitador.

El **CONCOR** es el módulo del IMPS encargado de la detección y corrección de errores. Tiene un propio lenguaje de programación con el cual el programador define cada una de las

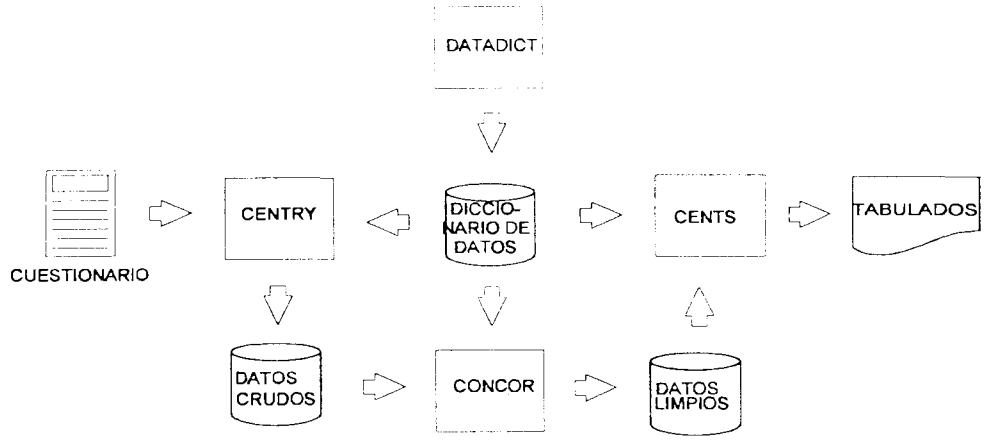


Figura 2.1: Principales módulos del IMPS

pautas de consistencia y si es del caso las reglas de corrección. Permite también obtener cifras sobre las frecuencias de los errores y las correcciones realizadas.

Para la obtención de los resultados el IMPS dispone del **CENTS** que utiliza la definición de las tablas de salida hecha por el usuario para procesar los datos limpios producto del programa anterior.

Otros módulos del sistema y que no se presentan en el esquema son el **QUICKTAB** que permite producir de manera rápida distribuciones de frecuencias y cruces simples de

variables. También existen otros programas dentro del IMPS como el **CENTRACK** que sirve para el control de los lotes de boletas, etc; el

CENVAR es un paquete para el cálculo de la varianza, muy útil para investigaciones muestrales. Además se dispone de un programa para la diseminación de resultados llamado **TRS** que sirve para la selección, recuperación e impresión de tablas estadísticas.

2.1.3 Organización de Datos

La estructura lógica de los datos es definida en el IMPS por medio del programa DATADICT que maneja el diccionario de datos.

El elemento principal en esta definición es el concepto de cuestionario que se asocia con la unidad de recolección. Este cuestionario está identificador por una llave única y se divide en tipos de registro. Cada tipo de registro corresponde con una entidad. Por ejemplo, en un Censo de Vivienda y Población, el cuestionario podrá tener dos tipos de registro, el uno corresponde a la vivienda y el dos a la población.

Cada tipo de registro contiene variables que tienen las características básicas de nombre, tipo, posición, tamaño y valores válidos.

Físicamente los datos se organizan de manera secuencial. Los datos de cada cuestionario conforman registros, por lo que, siguiendo el ejemplo anterior, primero se encuentra el

registro (1) de la vivienda y luego los registros (2) de cada persona de esa vivienda. Cada registro está identificado con el tipo de registro y la llave propia del cuestionario. Las variables se encuentran en posiciones fijas dentro de cada tipo de registro. El orden de los cuestionarios depende del orden en que se captan (o agregan) los datos. No se dispone de facilidades de compresión de datos, y estos se almacenan en ASCII por lo que es posible visualizarlos fácilmente con cualquier editor o programa similar.

2.1.4 Descripción de los principales módulos

2.1.4.1 Entrada de Datos (CENTRY)

El CENTRY dispone de dos modalidades de funcionamiento, una en que se desarrolla la aplicación de entrada de datos y la otra en que se procesa dicha aplicación.

Para la definición de la aplicación se cuenta con un concepto adicional a la jerarquía cuestionario-registro-variable que es el concepto de "batch" o lote de cuestionarios. Entonces, la aplicación consiste en una pantalla para las variables propias del lote, otra pantalla para el cuestionario y una (o más) pantallas para las variables de cada tipo de registro.

Durante la captura de datos, el digitador se encicla dentro de cada nivel de la jerarquía hasta completarla. Por ejemplo, primero se capta las variables propias del lote, luego la

del cuestionario, posteriormente la primera vivienda de ese cuestionario; para captar los datos de cada una de las personas de esa vivienda. Luego se capta la siguiente vivienda del cuestionario y así sucesivamente. Esta lógica predefinida puede ser alterada de acuerdo a las necesidades del operador.

A continuación se citan las características básicas del CENTRY:

- Revisión de la identificación del lote contra una tabla externa
- Revisión del dato captado contra el dominio de la variable
- Posibilidad de aplicación de chequeos de consistencia y de estructura
- Manejo de los saltos y del orden de captura de los datos
- Registro de estadísticas de rendimiento del digitador
- Verificación de los digitado

2.1.4.2 Consistencia y Corrección (CONCOR)

En el IMPS se integró el CONCOR, un programa desarrollado en los años setenta y que se ha utilizado para la depuración de datos de muchos censos y encuestas.

Una idea simplificada del funcionamiento de este programa es algo así como:

- se leen todos los datos de un cuestionario
- se aplican las pautas de consistencia contra ese cuestionario
- se reportan los errores encontrados, o
- se imputan valores correctos de acuerdo a las reglas de corrección
- se emiten las estadísticas de errores detectados e imputaciones realizadas para todos los cuestionarios

Para tal efecto, el CONCOR se apoya en el diccionario de datos del IMPS y a su vez dispone de un propio lenguaje de programación que permite al informático especificar tanto las pautas de consistencias como las reglas de corrección.

El CONCOR tiene una lógica predefinida que debe ser utilizada por el informático para establecer las revisiones necesarias. Esta lógica se compone de lo siguiente:

- un ciclo de lectura general que recorre secuencialmente todo el conjunto de datos, manteniendo en memoria los datos completos del cuestionario que se está procesando

- ejecución de la sección de código correspondiente al inicio del proceso de un nuevo cuestionario (AT-NEW-QUEST)

- ejecución del conjunto de instrucciones correspondientes a cada tipo de registro (FOR-EACH)

- ejecución de las instrucciones de final del cuestionario (AT-END-QUEST)

Una instrucción de CONCOR que detecta y reporta un error puede ser como:

```
If SEXO < > 1:2
        msg "SEXO ES INVALIDO" DUMP SEXO
end-if
```

en donde se revisa que el sexo este en el rango de 1 a 2 y si no es así se emite un mensaje de error con el valor de la variable SEXO.

2.1.4.3 Tabulaciones (CENTS)

El componente del IMPS encargado de emitir las tabulaciones es el CENTS cuya filosofía de funcionamiento se describe en el apartado de Producción de Resultados de esta guía.

El informático debe especificar la construcción de cada tabla en varias instancias distintas que se describen a continuación y que son las entradas a las fases del proceso que permiten producir las tablas:

CENTAL

En esta primera fase, se debe detallar para cada una de las tablas el contenido de las celdas, como por ejemplo:

```
tally TAB01(SEXO,EDAD)
```

Entonces, el CENTS lee los datos y de acuerdo a su definición en el diccionario de datos acumula en el contenido de la tabla correspondiente. El resultado de esta fase es un conjunto de los llamados "tally blocks" para cada tabla definida.

CENCON

La fase CENCON permite al usuario especificar los niveles de agregación que se tendrán para cada tabla. La fase anterior produjo "tally blocks" para los niveles inferiores y aquí se consolidan hacia arriba obteniendo nuevos "tally blocks" para los nuevos niveles de cada tabla.

CENAREA

Esta fase se dedica a definir los nombres de cada uno de los cortes para los cuales se emitirá una tabla.

CENPREP

En esta última fase se realiza tanto el diseño físico de la tabla de salida incluyendo los títulos, pie de página, y todos los textos y la distribución de los campos, como también se define los procedimientos para el cálculo de las entradas dependientes de la tabla.

Para lo anterior el CENTS dispone de un "editor" que permite diseñar interactivamente el formato de cada tabla y las correspondientes instrucciones de cálculo de las entradas a la tabla que lo requieran.

2.1.4.4 Recuperación de tablas (TRS)

En las últimas versiones del IMPS se ha incorporado un módulo adicional para la recuperación de tablas. Este módulo llamado "Table Retrieval System" (TRS) es un sencillo ejemplo de la diseminación de información censal agregada.

La idea de este programa es mantener una gran cantidad de tablas almacenadas en el computador y dar al usuario las herramientas necesarias tanto para su organización como posteriormente para su recuperación.

Estas tablas se organizan de acuerdo a los áreas geográficas para las cuales fueron emitidas y clasificadas por las variables que contiene. Entonces, el usuario puede seleccionar una área geográfica determinada y recuperar las tablas pre-elaboradas para las variables que sean de su interés.

Este módulo puede ser una alternativa de diseminación de los resultados de investigaciones, ya que pueden usarse los CD-ROM o incluso un sitio en internet en vez de la publicación de grandes documentos impresos.

2.1.5 Otras consideraciones técnicas

El IMPS ha sido diseñado para que funcione en un ambiente DOS en computadoras PC-Compatible y requiere cerca de 20Mb de espacio en disco duro y no más de 640Kb de memoria convencional. Debido a lo anterior, el limitante principal se refiere al volumen propio de los datos a procesar.

Una característica importante de este paquete es que los módulos CONCOR y CENTS requieren de un compilador COBOL para su funcionamiento. Esto es debido a que estos programas generan un programa fuente sintácticamente correcto que debe ser compilado, para luego ejecutarlo.

2.2 REDATAM+

2.2.1 Introducción

El REDATAM+, cuyo nombre viene de **RE**cuperación de **DAT**os de **A**reas pequeñas por **M**icrocomputador, es un sistema que permite el acceso a combinaciones de grandes bases de datos, organizadas jerárquicamente, incluyendo datos de censos nacionales, estadísticas agregadas y archivos de grandes encuestas. Además, organiza y almacena todos los datos existentes en grandes archivos en la limitada capacidad de un microcomputador de tal forma que el usuario pueda obtener fácilmente tabulaciones u otro tipo de estadísticas para las áreas jerárquicas definidas [CELADE-91].

2.2.2 Conceptos Básicos

REDATAM+ no es un Sistema de Administración de Bases de Datos de uso general, por el contrario está diseñado para ser utilizado en el procesamiento de datos de censos y encuestas, sin embargo es posible su uso en cualquier actividad que se adapte a sus principales características técnicas.

La información dentro de una base de datos REDATAM+ se organiza estrictamente de manera jerárquica. Esto se debe principalmente a que los países se organizan

internamente en niveles jerárquicos. Entonces, el acceso a la información se realiza para universos constituidos por ocurrencias de estas áreas o conjuntos de áreas.

Los procesos sobre las bases de datos se efectúan para determinadas *áreas pequeñas* y no para la totalidad de la información. Es evidente que tratar de obtener un resultado para todos los casos de un censo en un microcomputador demorará bastante tiempo. Sin embargo estas *áreas pequeñas* pueden comprender conjuntos importantes de información.

La información censal, una vez depurada, se puede catalogar como estática debido principalmente a que un censo es una especie de fotografía de una población en momento dado, por lo que no está sujeta a actualización. Entonces REDATAM+ permite "generar" una base de datos para la cual no habrá posibilidad de modificar su contenido posteriormente.

La lógica del REDATAM+ en el uso de la información de una base de datos comienza con la preparación del **conjunto de selección** que se desea analizar, posteriormente es posible escribir un *programa* para la obtención de los resultados para ese conjunto. El lenguaje propio de REDATAM+ es muy similar a algunos paquetes estadísticos como el SPSS[12] y permite dentro de otras cosas construir y recodificar variables y emitir distribuciones de frecuencias, tabulaciones cruzadas y tablas listas para impresión. Debido a que REDATAM+ tiene un ámbito de acción definido claramente, se le ha

---

[12]Statistics Package for the Social Sciences. Paquete estadístico de uso general de gran utilización a nivel mundial.

provisto de interfases por medio de resultados o subconjunto de datos con otros paquetes como el SPSS, Lotus 123, y varios Sistemas de Información Geográfica (SIG) como el Arcinfo.

La estructura jerárquica de una base de datos REDATAM+ está organizada en **ramas** y **entidades**. Cada entidad representa un nivel determinado en una rama. Por ejemplo, un país hipotético llamado Nueva Miranda y que REDATAM+ utiliza de ejemplo, tiene una estructura que se presenta en la figura 2.2.

Nótese que la base de datos está contenida en una sola rama y que cada entidad representa un nivel jerárquico en la estructura del país.

Las cuatro entidades superiores COMUNA, DISTRITO, AREA y SEGMENTO están representadas con un recuadro doble, esto se hace para diferenciarlas del resto de entidades debido que son **identificables** o seleccionables. Las virtud de que una entidad sea seleccionable la define el usuario a la hora de construir la base de datos y permitirá que las ocurrencias de esa entidad participen en conjuntos de selección (áreas pequeñas). Por ejemplo, es posible crear una selección que comprenda toda la comuna de Puerto Nuevo, y algunos distritos de otra comuna. No es posible seleccionar ocurrencias de las entidades VIVIENDA y POBLACION.

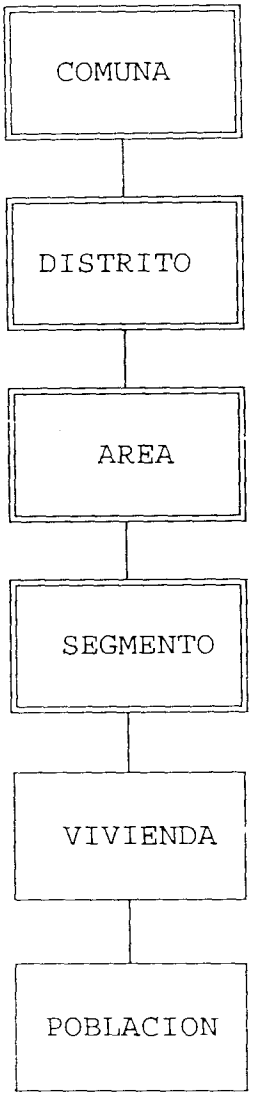


Figura 2.2: Estructura de la base de datos REDATAM+ del Censo de Población y Vivienda de 1982 de Nueva Miranda

Cada entidad esta compuesta por **elementos** o atributos que representan las variables, por ejemplo TIPO_PISO es un elemento de la entidad VIVIENDA. Las entidades seleccionables debe tener una variable "llave" que las identifique y en algunos casos

tendrán una variable con un "nombre jerárquico" que facilitará la labor de seleccionar los nombres.

2.2.3 Organización Funcional

El REDATAM Plus está organizado en cinco módulos los cuales le permiten al usuario, de manera interactiva, efectuar sus análisis. A continuación se describen brevemente cada uno de estos módulos.

CONFIGURACIÓN: Este módulo se utiliza para la definición de algunos parámetros necesarios para el REDATAM+. Aquí se indica la base de datos activa para la sesión, el idioma que se utilizará. También se dispone de opciones que permiten manejar las claves de protección y las opciones de omisión.

PROCESAMIENTO: Por medio de este módulo se producen los resultados. Su opción principal es el "Procesador Estadístico" que es un ambiente de desarrollo y ejecución de programas de acceso a la base de datos. Además es posible modificar las salidas sin volver a leer los datos.

SELECCIÓN: Todo proceso a realizar en el "Procesador Estadístico" debe indicar el subconjunto de datos a incluir. Para esto se tiene el módulo de selección que da la

facilidad al usuario para que "amigablemente" marque las ocurrencias de las entidades jerárquicas seleccionables que desea incluir.

DICCIONARIO: Este módulo es el encargado, como su nombre lo indica, de construir y dar mantenimiento al diccionario de datos. Con la opción de "Edición" es posible definir la estructura de la base de datos, las cualidades de las entidades y todo lo relacionado con las variables.

GENERACIÓN: La generación de bases de datos REDATAM+ es algo más complejo que su uso, sin embargo se tiene este módulo que permite al usuario efectuar la "carga" de la base de datos. Además se dispone de opciones para extraer subconjuntos de datos y analizar la integridad de la información almacenada.

Se distingue a dos tipos de usuario. El primero que es el encargado de construir el diccionario de datos y realizar la generación de la base de datos, que debe tener bastante experiencia en el uso de microcomputadores y procesamiento de datos estadísticos. Y un segundo usuario que tendrá acceso a la base de datos ya creada por medio de los módulo de "Selección" y "Procesamiento" para la producción de resultados.

2.2.4 Estructura interna

Una base de datos REDATAM+ (además del diccionario) está compuesta físicamente por dos instancias diferentes que se almacenan como archivos separados; estas son los archivos de datos que contienen la información de la ocurrencia de cada variable, y los punteros que indican las ocurrencias que tiene cada caso en su nivel inferior.

Archivos de Datos

La información se almacena en REDATAM+ en tablas invertidas, un archivo para cada variable. Por ejemplo, todas las ocurrencias de la variable PARENTESCO se encuentran en único archivo.

La utilización de esta estructura interna es una característica fundamental del REDATAM+.

Otra característica determinante del REDATAM+ es su facilidad de almacenar a nivel de bit. Esto quiere decir que la ocurrencia de un atributo estará registrada en un mínimo de espacio. Por ejemplo, se conoce que la variable ESTADO_CIVIL tiene como respuestas válidas los códigos 1 a 6 (soltero, casado, viudo, divorciado, unido, desconocido), en un archivo tradicional se necesita 1 byte completo (8 bits), por el contrario, en un archivo REDATAM+ esta variable ocuparía únicamente 3 bits.

Esta consideración, además de reducir el tiempo de acceso en la recuperación de datos, permite almacenar la información en una pequeña parte del espacio que se requeriría en otro sistema. Se habla de un 20% a un 25% de la cantidad de espacio original.
Las variables que identifican las ocurrencias de las entidades no pueden almacenarse de esta forma y tendrán que ser del tipo caracter.

Punteros

Como en la mayoría de las bases de datos jerárquicas, los información están físicamente organizada en el orden de la estructura jerárquica propia. En este caso, el orden de los tuples dentro de la tabla es determinante.

Para que los procesos que acceden a la base de datos encuentren los casos que correspondan a la selección hecha por el usuario es necesario identificar de donde a donde se encuentran las ocurrencias de determinada entidad. Para esto se utilizan punteros.

Cada entidad tiene un conjunto de punteros (un archivo físico) que indica cuantas ocurrencias existen para cada ocurrencia de la entidad superior. Obviamente, la entidad superior tiene un conjunto de punteros falso. Por ejemplo, si tomamos el puntero de DISTRITO (que une esta entidad con COMUNA) se tiene:

20 15 8 10 11 11 6[13]

lo que significa que la primera comuna tiene 20 distritos, la segunda 15 y así sucesivamente.

Entonces, cada puntero tiene el número de casos de la entidad superior . REDATAM+ emplea 4 bytes (un LONG) para almacenar cada ocurrencia del puntero.

De esta forma cuando se presenta un requerimiento para una selección determinada, el REDATAM+ "procesa los punteros" y obtiene únicamente los datos de los casos que concuerden con la escogencia hecha por el usuario.

2.2.5 Otra información técnica

El REDATAM+ funciona en microcomputadores IBM-compatibles y requiere una versión del sistema operativo DOS 3.1 o posterior. Prácticamente cualquier equipo con disco duro y al menos 640Kb de memoria es capaz de soportar al REDATAM+.
El programa se escribió en lenguaje C utilizando la biblioteca de soporte C-Worthy.

---

[13] En realidad este puntero guarda la cantidad acumulada, por lo que en realidad se vería así: 0 20 35 43 53 64 75

Ha sido desarrollado en CELADE en Santiago de Chile y se puede obtener copias a un costo entre US $ 75 y US $ 200 dependiendo del tipo de usuario.

3. CASIC[14]: Recolección de información asistida por computador

3.1 Introducción

Más de la mitad de los costos de producir estadísticas se encuentran en la recolección de los datos, la entrada, revisión, corrección y la codificación [Blom-94]. Esto ha motivado a que desde hace algún tiempo, se realicen esfuerzos en mejorar la eficiencia con que se llevan a cabo dichas actividades.

Los esfuerzos que se han hecho, se han centrado principalmente en incorporar la tecnología existente a estas actividades lo que ha llevado al concepto de CASIC.

Sin embargo, esto no es tan novedoso, ya que desde la década de los sesenta se comenzó a utilizar el teléfono para realizar encuestas de mercado y otras pequeñas investigaciones de opinión, en lo que ahora se conoce como "Computer Assisted Telephone Interviewing" (CATI). Posteriormente, luego de la aparición de las computadoras portátiles se desarrolló la idea de "Computer Assisted Personal Interviewing" (CATI) en la que cada entrevistador dispone de una computadora a la hora de realizar la entrevista [Connett-94].

---

[14] Computer Assisted Survey Information Collection

Según [Blom-94], CASIC es un sistema de obtención de información que usa la computadora como principal medio para la recolección de los datos del informante y para la transmisión de los datos a otros sitios para el procesamiento posterior.

Sin embargo, detrás de este concepto se encuentra el asunto trascendental que es realizar la captura de los datos (y algún procesamiento adicional) durante la ocurrencia del evento que genera la información. De esta forma se minimizará la ocurrencia de errores facilitando el procesamiento posterior de los datos, y mejorando la calidad de los resultados que se van a producir.

Dentro del concepto CASIC se encuentran tres grandes áreas: la entrevista asistida por computador (CAI), la auto-entrevista (CASI), y el apoyo del computador a la tradicional "Paper and Pencil Interviewing" (PAPI). Dentro de cada una de estas áreas existen técnicas con características particulares.

3.2 Entrevista asistida por computador (CAI)

En esta primera área del CASIC se encuentran las técnicas que se refieren a un entrevista realiza con el apoyo de un computador. Por lo tanto se identifica la participación del entrevistador y de un informante, pero el formulario se reemplaza por un programa de computador.

El computador controla la entrevista, determina el orden en que deben ir haciéndose las preguntas y revisa las respuestas tan pronto como son digitadas [Bethlehem-91]. Nótese que este programa se diferencia en gran manera del típico programa de entrada de datos, pues tiene un alto contenido de interactividad para obtener la respuesta adecuada del informante y corregir en la misma captura los errores que se presenten.

A continuación se comentan las técnicas más importantes en la entrevista asistida por computador:

3.2.1 Entrevista telefónica asistida por computador (CATI)

Esta es la técnica más antigua que se conoce en este tema y se basa en una llamada telefónica que el entrevistador realiza al informante. Ha sido ampliamente usada en investigaciones de mercado y encuestas de opinión. Sin embargo su principal inconveniente es que la muestra de la población que se puede incluir en la investigación está limitada por la tenencia de teléfono.

3.2.2 Entrevista personal asistida por computador (CAPI)

En este caso, el programa de captura de datos es cargado en computadores portátiles que llevan consigo los entrevistadores al campo. Esta técnica combina las ventajas de la entrevista cara a cara con las ventajas de la asistencia del computador.

Dependiendo de las características de la investigación (tamaño, plazo, etc.) los datos en la computadoras portátiles pueden ser enviados vía modem a las oficinas centrales para la agregación correspondiente.

3.3 Auto-entrevista (CASI)

Este grupo de técnicas, también llamadas "Computerized Self-Administered Questionnaires" (CSAQ), se caracterizan por la inexistencia del entrevistador, ya que el propio informante introduce los valores de sus respuestas al computador. Evidentemente, las técnicas que se discuten a continuación, tienen una aplicación real dependiendo de las características de la investigación a realizar.

3.3.1 Cuestionario electrónico

El cuestionario se transforma en este caso en un programa de computador orientado al informante. Este programa puede hacerse llegar al informante por varios medios. Puede ser copiado a un disquete y enviado por correo, o simplemente descargado desde alguna red.

El informante interactúa con el programa, en lo que sería la entrevista, y completa las respuestas introduciendo los datos al computador.

Esta técnica es útil principalmente en investigaciones en que el sujeto es una empresa, ya que probablemente tiene computadores en donde ejecutar el programa.

3.3.2 Respuestas por teléfono de tonos (TDE)

Esta técnica es bien conocida en ciertas actividades como en los servicios bancarios o las compras por correo. Se basan en la introducción de las respuestas por medio de los tonos que emite el teléfono al presionar los distintos dígitos.

Evidentemente, la utilidad de esta técnica es mayor en investigaciones en las que las preguntas y sus respuestas puedan ser limitadas a las características propias de esta tecnología.

3.4 Asistencia del computador en las técnicas tradicionales

Debido a los medios tradicionales de recolección de datos seguirán usándose, principalmente en los países subdesarrollados, se han incorporando avances tecnológicos a esas técnicas. Aquí se mantiene el concepto de entrevista en la que el entrevistador llena un cuestionario en presencia del informante que emite sus respuestas a las preguntas.

3.4.1 Lectura de caracteres ópticos (OCR)

La principal mejora en esta caso es el "scanning" de los cuestionarios en donde la imagen es trasladada al computador como un archivo gráfico. Posteriormente, este archivo es interpretado por el computador para conocer las respuestas presentes en ese cuestionario.

Esta técnica sustituye la digitación convencional de los datos del cuestionario, y es comentada en el apartado referente a la entrada automática de datos.

BIBLIOGRAFIA

Bethlehem, Jelke. *THE BLAISE SYSTEM: For integrated survey processing.* Netherlands Central Bureau of Statistics. march 1991.

Blom, Evert. *Building Integrated Systems of CASIC Technologies at Statistics Sweden.* 1994 Annual Research Conference and CASIC Technologies Interchange: Proceedings. USA. 1994

Bravo Cabria, 'M. SPEER & GEIS: Dos sistemas para la depuración de datos cuantitativos'. *Estadística Española.* núm 127, v. 33, 191-242. España. Mayo-Agosto 1991.

Chamberlain, K. Durding, B. *Editing of Population Census Data: How much is enough?.* International Statistical Programs Center. U.S. Bureau of the Census. USA. 1992.

CELADE, Naciones Unidas. 'REDATAM-Plus: Recuperación de Datos de Áreas Pequeñas por Microcomputador'. Santiago de Chile, 1991.

Connett, William E; *et al.CAI Systems -- The User Perspective.* 1994 Annual Research Conference and CASIC Technologies Interchange: Proceedings. USA. 1994

Cotton, P. 'PRODUCE: La tecnología de Base de Datos para el acceso eficiente a datos de población'. *Notas de Población*. n 22, 49-65. Santiago, CHILE. Abril 1980.

Cushing, J. *DHS Data Processing Strategy*. DHS Demographic and Helth Surveys. World Conference Vol II. USA. 1991.

Dekker, Arij. *CENSUS DATA PROCESSING AND DISSEMINATION*. United Nations, New York, USA. 1995.

Dekker, Arij. *Innovative Computer Methods for Demographics Enquiries and the Dissemination of Population Information*. 1996.

Di Battista, G. Batini, C. 'Design of Statistical Databases: A Methodologuy for the Conceptual Step'. *Inform. System*. v13, 407-422, n 4. USA. 1988.

Ellis, C. 'Population Census Data Processing in Latin America. New Developments'. *I.S.I. 48th Session*. Le Caire, 1991.

Muñoz, J. 'Data Management of Complex Socio-economic Surveys: from Questionnaire Design to Data Analysis'. *'I.S.I. 47th Session'*. Paris, 1989.

Naghi Namaktoroosh, M. *Metodología de Investigación*. México. LIMUSA. 1987.

NACIONES UNIDAS. *Desarrollo y diseño de los cuestionarios de Encuestas*. Programa para Desarrollar la Capacidad Nacional de Efectuar Encuestas de Hogares. Nueva York. 1986

NACIONES UNIDADES. *Manual de Procesamiento de Datos de Censos de Población con Microcomputadoras*. New York, 1991.

NACIONES UNIDAS. *Survey Data Processing*. National Household Survey Capability Programme. New York. 1982

NACIONES UNIDAS. *The use of Microcomputers for Censos Data Processing*. New York. USA. 1989.

Ortúzar, J. 'Revisión automática de datos de censos y encuestas mediante el uso de computadores medianos y pequeños'. *Notas de Población'*. n 17, 63-83. Santiago, CHILE. Agosto 1978.

Pierzchala, M. 'A Review of the State of the Art in Automated Data Editing and Imputation', *Journal of Official Statistics*. v 6, n 4. 355-377. Sweden. 1990.

Sierra Bravo, R. *Técnicas de Investigación Social. Teoría y Ejercicios*. España. PARANINFO. 1988.

Silva, A. *Procesamiento de Datos Censales: Lecciones de la Experiencia Latinoamericana*. CENTRO LATINOAMERICANO DE DEMOGRAFIA. Serie A N 170. Santiago de Chile, junio de 1985.

Toro, V. Chamberlain, K. *The use of Microcomputers for Census Processing in Developing Comuntries: An Update*. International Statistical Programs Center. U.S. Bureau of the Census. USA. 1988.

Trevor, C. *DHS Date Editing and Imputation*. DHS Demographic and Helth Surveys. World Conference Vol II. 1991.

U.S. Bureau of the Census. *Population Census Data Processing on Microcomputers*. USA. 1989.

U.S. Bureau of the Census. *1994 Annual Research Conference and CASIC Technologies Interchange: Proceedings*. USA. 1994

U.S. Bureau of the Census. *CONCOR (CONsistency and CORrection. User's Guide*. USA. 1995.

U.S. Bureau of the Census. *IMPS Integrated Microcomputer Processing System*. USA. 1995.

U.S. Bureau of the Census. *1990: Census of Population and Housing History*. Part C. USA. 1995.

Van de Pol, F. 'Data Editing of business survey: an overview'. *Netherlands Official Statistics*. v. 10, 32-37. winter 1995.

NOTA BIOGRÁFICA

Daniel Antich Montero (San José, 1956) es graduado en Ingeniería Industrial en la Universidad de Costa Rica en 1980, sin embargo su actividad laboral y profesional la ha dedicado exclusivamente a la computación. Ha trabajado con diversas empresas e instituciones nacionales e internacionales así también como consultor independiente, y su principal actividad ha estado relacionada en muchas ocasiones con el procesamiento de los datos de censos y encuestas. Actualmente trabaja para el Programa de Plaguicidas de la Universidad Nacional (PPUNA).