Organización de las Naciones Unidas para la Alimentación y la Agricultura

Taller Regional sobre las "Directrices Operativas" del Programa Mundial de Censos de Agricultura 2020 (CAM 2020) de la FAO

Santiago, Chile
22-26 de abril, 2019

Capítulo 21: Proceso y archivo
Capítulo 22: Acceso seguro a microdatos

Sesión 8: Post-enumeración

Miguel GALMÉS

CONTENIDO

Cap. 21: Procesamiento de datos

1. Introducción
2. Hardware y Software
3. Pruebas de programas de computadora
4. Actividades de procesamiento de datos
5. Edición de datos
6. Imputación
7. Validación y Tabulación de datos

Cap. 21: Archivo de datos

1. Introducción
2. Buenas prácticas para las Características de Integridad del Contenido Digital
3. Evaluación y auditoría

Cap. 22: Acceso seguro a los microdatos

1. Qué son los microdatos y por qué son útiles
2. Metadatos y divulgación Estadística
3. Tipos de acceso

Cap. 21a: Procesamiento de Datos

Introducción

- El procesamiento de datos incluye la codificación de datos, la entrada, la edición, la imputación, la validación y la tabulación.

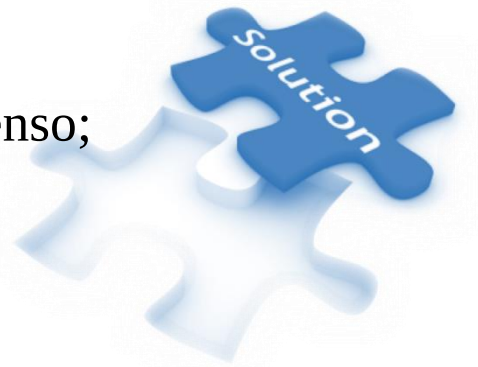
- El procesamiento de datos depende de la capacidad del país en términos de Tecnología de Información y Comunicación (TIC), es decir hardware, software e infraestructura, incluyendo decisiones sobre el método de recopilación de datos (ej. PAPI o DMC).

Hardware y Software

La **estrategia de TIC** para el censo debería formar parte de la estrategia general del censo agropecuario. Depende en gran medida de la opción de recopilación de datos y de la modalidad elegida para la toma del censo. En una etapa inicial se debe tomar la decisión para disponer del tiempo suficiente para probar e implementar el sistema de procesamiento de datos.

Se deben tratar las siguientes cuestiones claves de gestión:

- Orientaciones estratégicas para el programa del censo, a menudo relacionadas con duración y el costo;
- Infraestructura tecnológica existente;
- Nivel de soporte técnico disponible;
- Capacitación del personal de la agencia del censo;
- Tecnologías utilizadas en censos anteriores;
- Establecer la viabilidad de la tecnología;
- Coste-beneficio.



Hardware y Software– cont.

Requisitos de Hardware

Principales características del procesamiento de datos del censo agropecuario:

- Grandes cantidades de datos para ingresar en poco tiempo con usuarios múltiples y modo de procesamiento de servidores en paralelo;
- Se requieren grandes cantidades de almacenamiento de datos;
- Transacciones relativamente simples;
- Un número relativamente grande de tablas por preparar;
- Uso extensivo y simultáneo de archivos de datos en bruto;
- El método de capturar datos que ha elegido la oficina central del censo.

Equipo **de hardware básico**:

- Muchos dispositivos de captura de datos (PC, DMC, según el modo de recopilación);
- Un procesador/servidor central y redes;
- Impresoras gráficas rápidas y de alta resolución.

El **número** requerido de PC/ dispositivos de mano deben ser cuidadosamente considerado.

Software

- Debe permitir un procesamiento de los datos sin complicaciones;
- Debe utilizar un mantenimiento de software standard del fabricante (para permitir la portabilidad de los datos).

Prueba de programas de computadora

- Para escribir programas informáticos, se requiere un tiempo considerable.
- Los programas informáticos se deberían probar al verificar los resultados de la detección de errores y de las tabulaciones para un grupo de 100-500 cuestionarios (o menos si el personal calificado no está disponible).
- Los datos utilizados para tales pruebas se deben tabular manualmente para verificar cada ítem o su clasificación en las tabulaciones.
- Los censos piloto ofrecen una buena oportunidad para realizar pruebas finales y exhaustivas de sistemas y programas informáticos, incluyendo la transferencia de datos (en el caso de DMC, CATI, CASI).

Actividades de procesamiento de datos

Abarca:

1. Codificación y captura de datos
2. Edición de datos
3. Imputación
4. Validación y tabulación
5. Cálculo del error de muestreo y análisis de datos adicionales.

1. Codificación de datos y captura de datos

Codificación de datos: Es la operación en la que la información original del cuestionario en papel, tal como la registran los encuestadores, se reemplaza por un código numérico requerido para el procesamiento, puede ser manual y/o por computador.

Métodos de entrada de datos:

- Manual;
- Escaneo óptico:
  - Reconocimiento inteligente de caracteres (ICR);
  - Reconocimiento Óptico de Marcas (OMR);
- Dispositivos de mano (DMC);
- Internet y entrevistas telefónicas asistidas por computadora (CASI/CAWI y CATI).

Actividades de procesamiento de datos (cont.)

Entrada manual de datos

- Consume mucho tiempo
- Sujeta a errores humanos
- Se necesita más personal
- Se requiere un procedimiento de verificación rigurosa
- El software es simple de usar

Procesamiento de datos (cont.)

Entrada de datos con lector óptico

Solución ICR	Solución OMR
Ventajas: • Reducción de gastos por salarios (las respuestas se pueden codificar automáticamente); • Ahorros adicionales si se usan imágenes electrónicas en lugar de formularios físicos; • La codificación automática proporcionará mejoras en la calidad de los datos, porque garantiza un tratamiento uniforme de respuestas idénticas; • El tiempo de procesamiento se puede reducir; • El diseño del formulario no necesita ser tan estricto como el requerido para OMR; • Permite la presentación digital de formularios que resulta en la eficiencia de almacenamiento.	**Ventajas:** • La captura de respuestas en la casilla es mucho más rápida que la entrada manual; • El equipo es razonablemente económico; • Es relativamente simple de instalar y ejecutar; • Es una tecnología bien establecida que se ha utilizado durante varios años en muchos países.
Desventajas: • Mayores costos de equipo debido al sofisticado hardware y software requerido; • Sustitución de caracteres, que puede afectar la calidad; • La puesta a punto del motor y el proceso de reconocimiento para reconocer con precisión los caracteres es fundamental con compensaciones entre la calidad y el costo. • Las respuestas manuscritas se deben escribir en un área de respuesta restringida.	**Desventajas**: • OMR puede reconocer marcas hechas con un lápiz especial en números o letras pre-impresas en cuestionarios especiales. • Precisión requerida en el proceso de impresión de los cuestionarios • Restricciones sobre el tipo de papel y tinta que se puede usar; • Precisión requerida en el corte de hojas; • Restricciones en cuanto al diseño del formulario.

Actividades de procesamiento de datos (cont.)

Entrada de datos del dispositivo de mano

- Uso del DMC con cuestionarios electrónicos, captura de datos completados directamente por los encuestadores
  - Rentable
  - Permite la codificación y edición automática.
  - Permite patrones de saltos.
  - Se requiere una prueba exhaustiva de la aplicación de captura de datos:
    - Pruebas funcionales
    - Pruebas de usabilidad
    - Pruebas de transferencia de datos.

Entrada de datos CASI/CAWI y CATI

- Se administra generalmente con otros métodos
- Es similar a la recopilación de datos con DMC – se usan los formularios en línea que generalmente no es la versión descargable exacta del formulario en papel sino una aplicación que guía al encuestado a través del cuestionario.
- Es esencial comprobar los patrones de flujo y de saltos del formulario en línea.

Edición de datos

- El proceso implica la revisión y el ajuste de los datos recopilados en el censo.
- Objetivo: controlar la calidad de los datos recopilados
- Efecto de la edición de cuestionarios:
  - lograr coherencia en los datos y consistencia en las tabulaciones (en y entre las tablas)
  - detectar y verificar, corregir o eliminar valores atípicos.

Edición manual de datos (cuando se usa PAPI)

- Una función importante de la edición manual es verificar que los cuestionarios están completados: disminuir la falta de respuesta y verificar que identifiquen apropiadamente la explotación como una explotación agropecuaria que cumple con los requisitos mínimos.
- Debe comenzar tan pronto como sea posible, después de la recopilación de datos y cerca de la fuente de los datos
- Errores a menudo se deben a una escritura ilegible en cuestionarios en papel
- Tiene algunas ventajas: identificar cuestionarios en papel que se deben devolver para completar; ayuda a detectar una enumeración pobre.

Edición de datos – cont.

Edición automatizada de datos

- Correcciones electrónicas de datos digitales
- Edición eficiente para los censos, en términos de costos, recursos requeridos y tiempo de procesamiento
- Verificar la credibilidad general de los datos digitales con respecto a:
  - *datos faltantes,*
  - *pruebas de rango,*
  - *y consistencia lógica y/o numérica.*
- Dos maneras:
  - interactivamente en la etapa de captura de datos: los mensajes de error pueden aparecer de inmediato en la pantalla y/o pueden rechazar datos a menos que se corrijan; es útil en el caso de errores simples, como los errores de codificación, pero puede ralentizar mucho el proceso de captura de datos. Usado principalmente para descubrir errores en la captura de datos, mientras que los casos más difíciles, como la falta de respuesta, se dejan para una operación de edición automática por separado. Usado con los métodos de recopilación de datos DMC, CATI o CASI/CAWI
  - procesamiento por lotes: después de la captura de datos; consiste en una revisión de muchos cuestionarios en un lote. El resultado suele ser un archivo con mensajes de error. Este método se puede usar en todos los métodos de recopilación de datos.

Edición de datos – cont.

Edición automatizada de datos – cont.

Dos categorías de errores:

- **críticos** – deben corregirse; incluso podrían bloquear el procesamiento posterior o la captura de datos.
- **no críticos** – producen resultados no válidos o inconsistentes sin interrumpir el flujo de las fases de procesamiento; Se deben corregir todos los errores no críticos que sean posibles, para evitar la sobre-edición.

- **Edición de datos (detección de datos) se aplica a varios niveles:**
- A nivel de ítem, conocido como "comprobación de rango";
- A nivel de cuestionario (donde los controles se realizan a través de ítems relacionados del cuestionario);
- Jerárquico que implica verificar ítems en sub-cuestionarios relacionados.

3. Imputación

- Es el proceso de abordar las respuestas faltantes, inválidas o inconsistentes identificadas durante la edición en base al conocimiento disponible.
- Siempre que se utilice, se debe establecer un indicador.
- Se utilizan comúnmente dos técnicas de imputación:

  (a) imputación de plataforma fría (tablas de consulta estáticas)

  (b) imputación de plataforma caliente (tablas de consulta dinámicas).
- Se puede hacer de forma manual o automática por computadora.
- Considerar los siguientes aspectos para la edición automática y la imputación:
  - el objetivo inmediato en un censo agropecuario es recopilar datos de buena calidad. Si solo se descubren algunos errores, cualquier método se puede considerar satisfactorio para corregirlos;
  - es importante mantener un registro del número de errores descubiertos y la acción correctiva (por tipo de corrección);
  - la no respuesta siempre se puede tabular como tal en una columna separada.
  - la redundancia de la información recopilada en el cuestionario es útil para ayudar a detectar errores.

4. Validación de datos y tabulación

Validación

- Se debe realizar de forma paralela a los demás procesos.
- Todos los ítems de datos deben ser verificados por consistencia y exactitud para todas las categorías en los diferentes niveles de agregación geográfica.
- Macroedición - es el proceso de validación de datos a nivel agregado. El análisis se centra en aquellos errores que tienen un impacto en los datos publicados.
- La validación antes de que los datos salgan del centro de procesamiento garantiza que errores significativos se corrijan en el archivo final.
- Este archivo final se puede utilizar como la base de datos de origen para la producción de todos los productos de difusión.

Tabulación

- La tabulación es una parte importante del censo. Las tablas y los mapas son los resultados más visibles de toda la operación del censo y el producto más utilizado.

5. Cálculo del error de muestreo y análisis de datos adicionales

- Cuando se utiliza el muestreo, por ejemplo, para los módulos complementarios en el enfoque modular o los módulos de rotación en la modalidad integrada de censo y encuestas, los factores de expansión se deben calcular y aplicar de acuerdo con el diseño de muestreo.
- Los datos se pueden agregar usando fórmulas de estimación y factores de expansión apropiados.
- Los datos no se pueden utilizar y evaluar correctamente a menos que se asocie una indicación del error de muestreo con los valores obtenidos. En algunos casos, se introducen indicadores, advirtiendo al usuario del rango del error de muestreo, por tanto, de la calidad de los datos publicados.

Ejemplo de país: Certificación de datos de Canadá

- Statistics Canada tiene un comité de certificación (compuesto por responsables del censo y expertos en agricultura) que al final del proceso de validación revisa y certifica oficialmente los resultados. Cada variable del censo se revisa y certifica (por área geográfica). La información presentada al comité debe:
  - Anticipar los resultados del censo (pronóstico, otras encuestas, consultas con expertos de la industria).
  - Alinear los resultados con el contexto socioeconómico actual.
  - Comparar resultados con datos históricos, administrativos, de encuestas y otras variables correlacionadas.
  - Describir impacto del procesamiento/validación en los datos brutos.
  - Describir el impacto de los productores que dejaron la industria, los recién llegados y los grandes jugadores.
  - Recomendar al comité que los datos sean: • publicados; • publicados con advertencia; • diferidos para más investigación; o • no publicados.

CAP. 21b: ARCHIVO DE DATOS

(RLG – OCLC Report, RLG, Mountain View, CA May 2002)

Introducción

- El archivo del material del censo implica muchos aspectos, como la documentación técnica, los archivos de datos, los programas de TIC, etc.
- El objetivo es guardar los archivos de microdatos del censo.
- Al igual que otros datos, los datos del censo pueden tener un valor cultural e institucional para un futuro lejano. Por lo tanto, es importante tomar medidas adecuadas para proteger físicamente los datos.
- Es importante recalcar que haya copias de seguridad de los datos (en línea y fuera de línea).
- Este aspecto se debe integrar plenamente en el proceso de planificación y de presupuesto del censo para que se tomen medidas adecuadas con tiempo.
- Por suerte, las normas de preservación digital permiten que las oficinas censales administren datos digitales a largo plazo.

Buenas Prácticas para las Características de Integridad del Contenido Digital de la Red Internacional de Encuestas de Hogares (IHSN, 2009)

Característica de Integridad	Buena Práctica relacionada
Contenido: asegura que los elementos esenciales del contenido digital estén conservados	Se espera que una oficina de censo agropecuario identifique y activamente gestiones los datos a ser conservados.
Fijeza: requiere que los cambios para el contenido se graben perfectamente desde el momento de la creación hacia adelante.	Como mínimo, esta característica se podría dirigir a través del uso normal de una suma de controles para detectar los cambios intencionales o involuntarios de los datos, y notificar a los directores de datos para la acción.
Referencia: asegura que el contenido es única y específicamente identificable en relación con otro contenido a través del tiempo	Por ejemplo, se requiere una oficina de censo agropecuario para asumir y mantener un enfoque identificador persistente, que es un sistema para asignar y dirigir identificadores duraderos que permiten que se hagan referencias constantes y únicas a materiales digitales en el tiempo.
Procedencia: requiere que el contenido digital sea localizable desde su origen (desde el punto de creación) o, como mínimo, desde el depósito en un repositorio digital	Esta característica requiere que una oficina de censo agropecuario grabe la información (tomada como metadato) sobre la creación y la acción que han afectado el contenido desde su creación (por ejemplo, los datos depositados en un archivo, convertidos de un formato de archivo a otro).
Contexto: documenta y dirige las relaciones o el contenido digital	Una oficina de censo agropecuario que mantenga los datos debe documentar las relaciones entre su propio contenido digital y, si es posible, con los datos que gestionan otras organizaciones.

Evaluación y Auditoría en el archivo de datos

- **Una autoevaluación de la conservación** – demostrar el cumplimiento de los estándares establecidos y ayudar a las oficinas del censo a que "valoren sus elementos digitales con el propósito de que aborden las incumbencias de la preservación digital, teniendo en cuenta el alcance, las prioridades, los recursos, y el estado general de la preparación"[1].
- Después, las oficinas del censo pueden crear sus propias políticas de preservación, reflejando "el mandato de la organización para preservar datos".
- Están disponibles y se recomiendan las evaluaciones formales que incluyen:
  - Garantía de aprobación de datos;
  - Lista de control TDR - un riguroso estándar ISO (16363);
  - auditoría de repositorio digital basado en la evaluación de riesgo (DRAMBORA).
- Los "Diez principios para los repositorio de preservación digital" resumen los criterios principales para los repositorios fiables.

[1] IHSN 2009, Documento de Trabajo No.3: Principios y buenas prácticas para la preservación de datos

Capítulo 22: ACCESO SEGURO A LOS MICRODATOS

(RLG – OCLC Report, RLG, Mountain View, CA May 2002)

¿Qué son los microdatos y por qué son útiles?

- Los microdatos se refieren a la información que se graba por o del encuestado cuando se realiza una encuesta o censo.
- Para el censo agropecuario, esto correspondería a los datos que se han recopilado de la explotación.
- Cada fila en los microdatos corresponde a una explotación y cada columna a las variables de datos.
- Además de las etiquetas de las variables, los investigadores y otros usuarios de datos necesitan metadatos que les ayuden a comprender códigos, definiciones y conceptos en los que se apoyan los datos recopilados: *qué miden los datos y cómo se crearon, para comprender la calidad de los datos*.
- Los microdatos permiten a los investigadores utilizar los datos del censo o de la encuesta para analizar preguntas que requieren un análisis más preciso que las tabulaciones originales.

Microdatos agropecuarios

- Característica particular: las explotaciones entran en la definición de empresas o negocios porque son unidades de producción.
- Los riesgos y las técnicas de control de divulgación para los datos comerciales difieren de los de las encuestas de hogares: Los datos empresariales o comerciales de las fincas comerciales grandes también suelen corresponder a poblaciones pequeñas y por lo tanto, más difíciles de anonimizar.
- Sin embargo, se puede decir que las características de los *datos de censos y encuestas agropecuarias* comparten características con los datos de *encuestas de empresas* y *encuestas de hogares*. Esto se debe a que las explotaciones agropecuarias consisten, a menudo, en *pequeñas explotaciones* gestionadas por particulares, a menudo *de subsistencia*, y *fincas o empresas comerciales*, a menudo de gran escala y gestionadas por entidades jurídicas, que son más similares en características a las entidades en encuestas empresariales típicas.

Confidencialidad y marcos legales

Confidencialidad

- Consideraciones primordiales a la hora de elegir un sistema de acceso a microdatos son: satisfacer las necesidades de los investigadores y garantizar la mayor protección para mantener la privacidad de los informantes.
- El acceso a los microdatos requiere que las agencias estadísticas equilibren las demandas que emanan de la comunidad investigadora con sus requisitos legales, para mantener la confidencialidad de la información que han recopilado de los informantes.

Marco legal y político

- La agencia debe cumplir con el marco legal y los estatutos en que opera: mantener el apoyo de los informantes y al mismo tiempo, se reconoce que existen métodos para garantizar la confidencialidad y hacer que los microdatos estén disponibles para fines estadísticos.
- Es importante tener políticas claras sobre las acciones que la agencia puede tomar con respecto al acceso a los microdatos del censo, y que esta información esté disponible para el público de forma transparente.

Revelación de datos

- La protección de la confidencialidad significa que se deben realizar los mejores intentos para garantizar que el archivo no revele datos individuales.
- Una **revelación** tiene lugar cuando alguien que está usando un archivo de microdatos reconoce o aprende algo que no sabía antes sobre un encuestado en el censo o encuesta o cuando existe la posibilidad de que una explotación o persona en el hogar del productor sea re-identificado por un usuario usando la información contenida en el archivo.
- Hay tres formas principales en las que puede tener lugar la revelación:
  - **Revelación identificada:** cuando se deja en el archivo un identificador directo (por ejemplo, un nombre, número de teléfono o dirección) a partir del cual se puede conocer la identidad del encuestado.
  - **Revelación a partir de los atributos:** se produce si un atributo o una combinación de atributos (por ejemplo, una finca comercial grande o un tipo de cultivo raro) se pueden asociar directamente con un encuestado concreto. Las personas con conocimiento de la región probablemente identificarían a esa persona.
  - **Revelación residual** ocurre cuando las recuperaciones sucesivas de un archivo se pueden comparar (sustraer) para aislar el valor de un encuestado.

Metadatos y divulgación estadística

Control de Divulgación Estadística

- La divulgación estadística se refiere al proceso para garantizar que se cumplan los requisitos de confidencialidad que rigen el trabajo de la Oficina/Instituto Nacional de Estadística (ONE/INE), y que el riesgo de revelar información sobre el encuestado es mínimo (también se conoce como anonimización).
- El riesgo de control de la divulgación estadística es afectado por muchos factores, incluyendo:
  - La sensibilidad de los datos;
  - La existencia de fuentes exteriores de información que se pueden usar para intentar volver a identificar a los informantes, jugando con combinaciones de variables que pueden re-identificar a un encuestado;
  - La habilidad de combinar los datos difundidos con datos públicamente disponibles;
  - Si el archivo de microdatos procede de una encuesta por muestreo o de un censo de enumeración completa.

DISCLOSURE

Control de divulgación estadística

Los procedimientos de control de la divulgación estadística implican:

- **El retiro de identificadores directos** como nombres, direcciones, números de teléfono, ubicación detallada de las explotaciones agropecuarias.
- **El retiro de identificadores indirectos** como ubicación detallada de unidades agropecuarias. Esto incluye coordenadas geográficas, ubicación de segmentos de muestra, ubicaciones de parcela o ubicaciones de segmento, grabadas tanto como información de atributo o como parte de una muestra de marco de área.
- La **aplicación** de la técnica **de anonimización,** basada en el riesgo de revelación identificado. Aquí no se tratan los detalles técnicos especificados de las técnicas y el software para ayudar la anonimización, pero existe una amplia literatura a disposición.
- **La evaluación del archivo** por su utilidad y por la pérdida de información.

DISCLOSURE

Tipos de acceso a microdatos

1. **Archivos de Uso Público** (PUF): se someten a un riguroso proceso de control estadístico para que la posibilidad de identificar a los encuestados sea mínima.
2. **Archivos licenciados:** los archivos con licencia también se anonimizan, pero con la posibilidad de que se apliquen menos procedimientos de control.
3. **Instalaciones de Acceso Remoto (RAF):** una ventana de servicio proporcionada por las ONE permite a los investigadores suministrar el algoritmo y las herramientas que usarán. Personal del ONE ejecuta el trabajo y devuelve el resultado al usuario.
4. **Enclaves de datos:** una facilidad dentro del predio de la ONE a la cual los investigadores puede acudir para que les realicen su investigación a partir de los archivos suministrados a ellos.
5. **Empleado presunto:** el investigador es juramentado para trabajar con la ONE como un miembro temporal del personal.

Ejemplos de países

- Algunos países en la región como Bolivia, Costa Rica y Perú usan REDATAM para difundir en línea microdatos del censo de agricultura.
- REDATAM es el acrónimo de REcuperación de DATos para Áreas pequeñas por Microcomputador.
- REDATAM fue desarrollado por el Centro Latinoamericano y Caribeño de Demografía (CELADE), de la División de Población de la CEPAL.
- Costa Rica desarrolló una ***app*** para dispositivos móviles (Android) con el fin de difundir resultados del censo agropecuario en forma amigable y que incluía definiciones básicas para facilitar su comprensión.

Gracias