Documento de Referencia N° 6

Septiembre de 1989

ORIGINAL: ESPAÑOL

C E P A L
Comisión Económica para América Latina y el Caribe

Seminario Sistemas Computacionales para el Acceso de Usuarios a la Información Censal

Santiago, 12 al 15 de septiembre de 1989

LA CONFIDENCIALIDAD Y PRIVACIDAD DE LOS DATOS DEMOGRAFICOS CENSALES */

---

*/ Este documento fue preparado por Centro Latinoamericano de Demografía (CELADE), en el marco del Proyecto sobre Apoyo al Diseño y Preparación a la Ronda de los Censos del 90, administrado por la Comisión Económica para América Latina y el Caribe, CEPAL y financiado por el Fondo de las Naciones Unidas para actividades en materia de población.

LA CONFIDENCIALIDAD Y PRIVACIDAD DE LOS DATOS DEMOGRÁFICOS CENSALES

1. Algunos conceptos

La confidencialidad, tal como la vemos, significa en última instancia, la protección del secreto estadístico, o en otras palabras, impedir que algún registro o respuesta pueda ser individualizado. En el caso de los censos demográficos, se trata de impedir la identificación de una vivienda ó de una persona.

La importancia de mantener el secreto estadístico, que en la gran mayoría de los países de la región y del mundo es fuerza de ley, se basa en la propia confianza de los usuarios y del público en general en el sistema estadístico, porque, en la medida en que esta protección sea violada, y dicha violación sea conocida, el futuro de la calidad de la información estará en juego, ya que los informantes al saber que sus respuestas estarán disponibles e identificadas, no responderán de manera correcta los cuestionarios censales.

Tratándose de censos demográficos, la individualización no es tan crítica como en un censo económico (comercio, industria y servicios), ó como en un censo agropecuario. En éstos existe, por una parte, una gran diferencia entre tipos de establecimientos y, por otra, el número de ellos también es mucho menor.

En un censo demográfico, las preguntas son las mismas para todas las personas (salvo los cortes de edad y las preguntas de fecundidad), lo que no sucede con los censos económicos, donde hay bloques de preguntas específicas para cada rama de la industria, comercio, etc. Por eso, se espera que los registros de las viviendas y personas sean "más similares", ó mejor aún, "más uniformes" que los registros de un censo económico.

De esta manera, la privacidad afecta en mayor medida las viviendas y/o personas "menos parecidas", es decir, con diferentes características acentuadas, tales como el ingreso no común, la ocupación ó la rama de actividad económica singular, etc.

El problema de la identificación también se agudiza en razón inversamente proporcional al tamaño del universo de datos. Es decir, cuanto menor sea el número de casos, mayor deberán ser los cuidados con la confidencialidad.

La tendencia actual de los usuarios de los datos estadísticos va en dirección hacia una descentralización de la información, con un aumento de la planificación sectorial, decisiones a niveles subnacionales (regionales, municipales, etc.), lo que obliga a quienes generan los datos (oficinas nacionales de estadística) a prepararse para este tipo de demanda.

Esta tendencia hacia la descentralización provoca una mayor explotación de los datos a niveles geográficos menores. Por lo tanto, los cuestionarios censales deben ser identificados hasta los niveles geográficos menores (por ejemplo, manzana), con el objeto de posibilitar la recuperación de la información a estos mismos niveles, con lo cual se aumenta el riesgo del secreto estadístico.

La reducción del universo de estudio de un censo demográfico se ve afectada por tres elementos complementarios: a) nivel geográfico; b) selección de casos; y c) pulverización de celdas de resultados.

1.1 **Nivel geográfico.**

Cuanto menor sea el nivel geográfico en que se esté trabajando, mayor será la posibilidad de identificación de una vivienda ó persona. Por ejemplo, los resultados a nivel nacional están infinitamente menos sujetos a este problema que si tratamos a un sólo sector ó a un grupo de sectores censales.

1.2 **Selección de casos.**

Los "filtros" también funcionan en la misma dirección que los niveles geográficos, es decir, como se usan para restringir el universo de casos, cuanto mayor la restricción mayor la probabilidad de identificación. Como por ejemplo, al seleccionar las personas de un determinado grupo de edad, sexo y ocupación, se limita el universo de casos, aumenta la posibilidad de identificación.

1.3 **Pulverización de celdas.**

Por último, la extremada desagregación de los resultados también puede provocar el conflicto. Por ejemplo, al efectuar el cruce de grupo de edad por sexo, parentesco, estado civil y categorías muy desagregadas de ocupación, la matriz resultado puede ser muy diseminada, con celdas conteniendo uno ó dos casos.

Por otro lado, como en los censos demográficos no se ingresan los nombres de las personas (ni mucho menos sus apellidos), el secreto estadístico está "más protegido" que en otro tipo de encuestas.

2. La divulgación convencional

Por divulgación convencional se entiende las publicaciones, microfichas, microfilmes, etc., es decir, los medios no-magnéticos y, por lo tanto, pre-procesados con anterioridad por los institutos de estadística. En general, se trata de resultados agregados por región ó a nivel nacional y, por lo tanto, el riesgo de identificación es mínimo.

Además, esta preparación de los "productos" para entregar al "consumidor" exige un mínimo de análisis y de verificación ("control de calidad"), lo que aumenta la protección contra el "virus" de la identificación.

La "desidentificación" posee varias formas, desde las más drásticas (eliminar el cuadro), hasta las más sofisticadas (retocar los resultados por un proceso computacional de asignación de números aleatorios a las celdas problemáticas).

De esta forma, se considera que este tipo de información no pertenece al "grupo de riesgo".

3. Divulgación computacional

Por divulgación computacional, entendemos los "productos" grabados en medios magnéticos (discos, cintas) u ópticos (discos láser), que posibilitan al usuario un post-procesamiento ad hoc. Estos pueden ser clasificados en dos grupos: a) los productos agregados, y b) los desagregados.

3.1 Datos agregados

En este grupo está, por ejemplo, un diskette con todos los cuadros publicados, pero a nivel municipal. O el sistema SUPERMAP de la Space-Time Research, un paquete que accesa un disco CD-Rom con datos agregados espacialmente.

De la misma manera que en la divulgación convencional, estos productos exijen una preparación (y análisis) por parte de la oficina de estadística, antes de entregarlos al consumidor, y por ende, una posible "desidentificación" de los casos críticos.

Se considera, entonces, que este tipo de información tampoco pertenece al "grupo de riesgo".

3.2 Datos desagregados

Los datos desagregados (también llamados de microdatos), son exactamente aquellos recogidos en terreno, o sea, el contenido mismo de la boleta censal, con las respuestas a nivel de vivienda y persona.

Estos microdatos pueden ser clasificados también en dos grupos: centralizados y descentralizados.

3.2.1 Base de datos centralizada

Las oficinas de estadística con grandes computadores tienen la posibilidad de crear bases de datos con la información censal a nivel de microdato y, además, que esta información esté disponible para diversos usuarios simultáneamente a través de la utilización de terminales conectados con el computador central.

Estos terminales pueden ser "locales", es decir, instalados en la propia oficina de estadística, ó "remotos", instalados en otras oficinas (gubernamentales ó del sector privado).

En caso de no haber consultas "remotas" no existe el problema de identificación, ya que la información es consultada solamente por funcionarios de la oficina de estadística.

En el caso contrario, cuando hay acceso externo a las bases de datos, la posibilidad de una identificación de los registros no puede ser ignorada. Sin embargo, la "población de riesgo" es conocida, es decir, se sabe donde están instalados los terminales de acceso remoto y a través de señas de utilización ("passwords"), se puede controlar el acceso a la base de datos.

Por otro lado, el control de esta población de usuarios externos tiene que ser más sofisticado, no limitándose a una decisión binaria entre "si, se puede todo" y "no, no se puede nada". Si no se puede leer la base de datos, ¿para qué se necesita entonces el terminal? ¿Los datos no deben estar disponibles al público? (y otros argumentos similares).

Claro está que para implementar un sistema sofisticado de control, éste debería ser parte del mecanismo de acceso a los datos. Es decir, los archivos de datos censales no pueden ser almacenados de manera común y corriente y que cualquier programa (en cualquier lenguaje de programación) pueda leerlo. Estos archivos deben ser parte de una verdadera base de datos, lo que exije un paquete exclusivo de acceso a la misma, paquete éste que, dentro de otras características necesarias, inherentes al procesamiento estadístico, debe poseer también la facilidad de protección de la confidencialidad de los datos.

Como se vió en el capítulo 1, la definición de secreto estadístico no es muy precisa (el concepto sí es claro), y su implementación entonces depende de la interpretación de cada uno.

3.2.2 **Base de datos descentralizada**

Las oficinas de estadística entregan los archivos con los datos censales y éstos son utilizados directamente en los equipos de los usuarios.

Este es el "grupo de alto riesgo" en materia de confidencialidad de la información. El "presunto violador" tiene los datos y el equipo a su disposición y, aparentemente, ningún tipo de control a la vista!!

La situación es similar al acceso centralizado. Si estos archivos entregados al público son del tipo común, sin ninguna especie de formateo y pueden ser leídos por cualquier programa ó lenguaje, ellos estarían "desprotegidos". Si son del tipo base de datos, con un paquete específico de lectura, los requisitos para este paquete son los mismos ya descritos para una base de datos centralizada.

4. Protección de la información

Para la divulgación convencional, los sistemas de protección también pueden ser llamados convencionales, es decir, análisis de resultados antes de entregarlos, "desidentificación" de celdas poco pobladas, etc., procedimientos mencionados en el ítem 2, largamente utilizados en el pasado y conocidos de todos los institutos de estadística.

Para la distribución de datos agregados, además del mismo sistema de protección usado para la divulgación convencional, se adopta también la elección adecuada del nivel geográfico mínimo de agregación, nivel éste que asegura la total protección de los datos. Por ejemplo, no se entregan datos agregados a niveles menores que municipio (distritos, sectores, ó manzanas). Por supuesto que esta decisión limita el nivel geográfico de análisis a los usuarios, impidiendo que se hagan estudios a nivel distrital.

Para las bases de datos desagregadas (microdatos) de uso público, tanto centralizadas como descentralizadas, también se puede usar el criterio de limitación del nivel geográfico mínimo, eliminándose de la base de datos las variables geográficas de menores niveles (sector, manzana, etc.), con la misma limitación ya mencionada. Otras formas de protección dependen del paquete utilizado para el acceso a la base, como "passwords" privilegiadas, número mínimo de casos, "desidentificación" lógica de resultados controlados por el programa, etc.

Otra manera es eliminar las variables "conflictivas", como, por ejemplo, las variables de ingreso de la persona.

5. Comentarios

En el pasado, con los usuarios restringidos a la utilización de datos publicados por las oficinas de estadística, el concepto de confidencialidad de la información demográfica censal no era puesto en peligro, su aplicación era de manera indirecta, puesto que los resultados publicados (ó entregados en forma convencional) sufrían rigurosos (?) análisis antes de su entrega al público.

Con la masificación de la tecnología computacional, agregada a la tendencia a la descentralización de la planificación a niveles subregionales, los productos censales tienden a ser magnetizados, y cada vez más los usuarios necesitan tener acceso a los microdatos.

Las consecuencias de esta tendencia, analizadas optimísticamente, tienen varias ventajas, tales como la disminución de las publicaciones en papel, utilización generalizada de la información, con el consiguiente aumento del interés en los datos censales, etc.

Sin embargo, el hecho de "entregar" a los usuarios los microdatos censales acarrea implícitamente una posible violación del secreto estadístico, que puede ser evitado (ó disminuído) con las precauciones mencionadas.

De todas maneras, no se debería usar el criterio de la confidencialidad como una excusa o barrera para no divulgar los microdatos censales, especialmente porque, tratándose de informaciones demográficas (y no de censos económicos o agrícolas), el riesgo es mucho menor.

ref:\notas\89\8909semi.doc