TÉCNICAS DE PROTECCIÓN
Y SEGURIDAD DE DATOS ESTADÍSTICOS

Marta Más

EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA

Donostia-San Sebastián, 1
01010 VITORIA-GASTEIZ
Tel.: 945 01 75 00
Fax.: 945 01 75 01
E-mail: eustat@eustat.es
www.eustat.es

Indice

Capítulo

1

Introducción

Introducción y Objetivos

El objeto principal de este cuaderno persigue la clasificación y descripción de las técnicas de seguridad y protección de datos estadísticos, más importantes y comúnmente aplicadas por las agencias e institutos de estadística.

La necesidad de controlar el enorme flujo de datos e información que circulan en nuestros días por las grandes redes de comunicación, justifica la aplicación de estas técnicas sobre los datos, antes de que éstos sean publicados o difundidos. De esta forma se pretende evitar que el uso inadecuado de toda esta información, revierta en un daño o perjuicio sobre individuos o entidades, de los cuáles se tienen datos que son accesibles por cualquier usuario desde cualquier parte del mundo.

No se pretende, en ningún caso, vetar el derecho a la información que todos tenemos y que está fuera de toda duda, sino equilibrar éste con otro derecho básico que contempla la preservación de la intimidad del individuo. El objetivo primordial a la hora de aplicar estos métodos y técnicas consistirá en aportar una máxima calidad de información de la forma más segura.

Conceptos Básicos de Seguridad y Confidencialidad

Como en muchas áreas de conocimiento, la seguridad y protección de datos utiliza una terminología propia que debe ser entendida siempre dentro de su contexto. Es por eso necesario aclarar y definir previamente, conceptos que van a ser utilizados con posterioridad y de forma recurrente a lo largo de este cuaderno.

Definiciones Previas

- ***Dato.*** Representación de la realidad, el mundo, los individuos y las posibles relaciones entre ellos. En estadística, toda aquella información organizada para el análisis y la toma de decisiones.

- ***Dato Confidencial.*** Aquel que por motivos éticos, morales o adquiridos de común acuerdo con el encuestado o con el propietario de los datos, no puede difundirse o publicarse para su análisis o uso externo.

- ***Dato Sensitivo.*** Es aquel que, no siendo de carácter confidencial, permite sin embargo, estimar estrechamente o deducir información confidencial exacta. (*)

(*) En muchas ocasiones a lo largo de la redacción del cuaderno se van a utilizar indistintamente los términos confidencial y sensitivo, para referirnos a datos que no pueden ser publicados o cuyo conocimiento supone un riesgo para la seguridad de información confidencial.

- ***Dato Seguro.*** Es aquel que no es confidencial y además no aporta información sobre ningún dato que lo sea. Hablaremos de *conjuntos de datos seguros* o *tablas seguras* según el caso.

- ***Intruso.*** También denominado *espía de datos* o *atacante*, hace referencia a la persona o grupo de personas que, de forma individual o colectiva y de manera intencionada o no, comprometen información confidencial o sensitiva dentro de un sistema de información.

Intimidad y Confidencialidad

Para entender la necesidad de desarrollar técnicas específicas para el control de la divulgación de datos estadísticos, debemos partir del derecho básico de la persona a *la intimidad*. Entendemos por intimidad, *la libertad del individuo a decidir cuánto sobre sí mismo quiere que sea revelado a otros, cuándo y a quiénes*. Se trata por tanto de un derecho a la "propiedad personal" y puede ser entendido como un estado del individuo.

Paralelamente a éste, surge el concepto de *confidencialidad,* como el deseo de un determinado individuo por *restringir el acceso a información personal que puede ser utilizada para determinados fines o propósitos*. En contraste con la anterior definición, la confidencialidad puede ser vista como un estado de la información y los datos y no del individuo en sí mismo.(*)

Ambas acepciones se unen en una relación causa-efecto en dónde la revelación de información confidencial sobre un individuo puede considerarse como una invasión de su intimidad y por lo tanto como la transgresión de uno de sus derechos fundamentales.

En el ámbito legal, tanto en la estadística como en cualquier otra ciencia relacionada con temas humanos, la preservación de la confidencialidad viene regulada por normas específicas, que en muchos casos varían de un país a otro. El cumplimiento del denominado *secreto estadístico* viene contemplado a nivel estatal en la Ley de la Función Estadística Pública [24] y de forma más generalizada en las normas de EUROSTAT 1993.

Identificación y Divulgación

Definir lo que se entiende como *divulgación* de datos estadísticos confidenciales, no es una tarea fácil. Diversos autores difieren en la literatura sobre el tema a la hora de considerar cuándo se ha producido realmente una difusión de información confidencial. ¿Basta con identificar a un individuo concreto dentro de una población o es necesario conocer algún atributo confidencial del mismo? [19].

La *identificación* o pérdida del anonimato se produce al reconocer un conjunto de características o atributos (confidenciales o no) como pertenecientes a un determinado individuo de la población. En muchas ocasiones, compromete por sí sola la integridad de los datos y pone en entredicho la seguridad del sistema de información. Sin embargo, en muchos casos se requiere, a su vez, el conocimiento de información confidencial sobre la unidad o individuo de la población identificado, para considerar que se ha producido una verdadera divulgación. De esta manera y formalizando la definición, podremos decir que se produce una *difusión o divulgación* de carácter confidencial cuando:

(*) Rieken, H.V. (1983) "Solutions to Ethical and Legal Problems in Social Research" New York:Academic press 1-9.

i) Tiene lugar una identificación *(divulgación de identidad)*

ii) Se conoce información sensitiva o confidencial del individuo o unidad identificado de la población *(divulgación de atributo)*

La anterior definición nos lleva a considerar distintos tipos de divulgación, según la información que se conoce, entre los que destacamos:

- ***Divulgación Exacta.*** Se produce cuando se conoce el valor exacto del atributo confidencial que pertenece a la unidad o individuo de la población que ha sido identificado.

- ***Divulgación Estadística o de Intervalo.*** Se produce cuando se puede deducir un intervalo "muy estrecho" o una estimación muy aproximada para el valor exacto del atributo o característica confidencial.

- ***Divulgación Verdadera.*** Se produce cuando el valor exacto o estadístico conocido se corresponde con el valor real de la característica o atributo en el momento en el que se produce la difusión.

- ***Divulgación Aparente.*** Se produce cuando el valor exacto o estadístico conocido representa un valor posible o factible de la característica o atributo confidencial pero no es el válido o real en el momento de su difusión.

- ***Divulgación Positiva o Interna.*** La información que se conoce pertenece a la unidad o individuo identificado.

- ***Divulgación Negativa o Externa.*** La información que se conoce no pertenece a la unidad o individuo identificado sino a aquellos que lo complementan, es decir, se conocen características o atributos que no tiene el individuo en sí mismo pero sí otros directamente relacionados con él (p.ej. empresas competidoras, compañeros de profesión, etc...).

Todos los esfuerzos irán encaminados a evitar cualquiera de los tipos de divulgación existentes. No se trata de una tarea fácil, debido a las múltiples vías de acceso a la información que existen y a la diversidad de formatos en los que se publican los datos en la actualidad. Limitar el acceso en sistemas de información mediante claves y permisos especiales a determinados usuarios, ayuda a evitar divulgaciones exactas de datos confidenciales. Sin embargo se pueden producir ataques más indirectos que permitan deducir información confidencial a partir de datos aparentemente "inofensivos". Este es el denominado *Problema Inferencial* [18].

El Problema Inferencial

Como ya hemos explicado en el apartado anterior, una divulgación de carácter confidencial se produce al conocer atributos secretos o confidenciales de un determinado individuo o unidad de la población previamente identificado. Si la aplicación de las normas sobre el secreto estadístico y la preservación de la confidencialidad, garantizan que estos datos no están disponibles o no son accesibles para el público en general, ¿de qué forma pueden llegar a conocerse?.

Cómo evitar la inferencia o deducción de datos sensitivos o confidenciales a partir de información que aparentemente no lo es, es uno de los principales problemas a la hora de implementar técnicas eficientes de protección de datos. Algunos factores de

importancia a tener en cuenta a la hora de buscar soluciones al Problema Inferencial son los siguientes:

- *Variables Geográficas*. La información geográfica detallada puede dar lugar a la identificación de pequeñas áreas, donde es posible reconocer a individuos o unidades de la población como pertenecientes a ellas. En general, no suelen publicarse identificadores regionales directos (p.ej. códigos de regiones o áreas concretas) y algunas agencias y oficinas estadísticas limitan la publicación de información geográfica a zonas con un tamaño suficientemente grande como para que estas identificaciones no se produzcan.

- *Pesos muestrales.* Normalmente aplicados para corregir defectos de la muestra, aportan en muchas ocasiones información sobre los estratos de la población, de forma que un individuo que sea identificado como perteneciente a dicho estrato, poseerá las características que lo definen. Estos pesos hacen referencia en muchos casos a información regional o geográfica que no es posible difundir. La divulgación de estas características sitúa a la agencia o instituto que publica dichos datos, en una posición comprometida respecto a la preservación de la confidencialidad.

- *Información externa disponible*. El conocimiento previo que cualquier usuario puede tener sobre la población, puede ayudar a deducir información confidencial a partir de datos que inicialmente son seguros.

- La deducción de los *pasos a seguir* por un hipotético intruso para llevar a cabo una inferencia de carácter confidencial, ayudarán a implementar técnicas que eviten que ésta se produzca.

Distintas soluciones basadas en estos dos últimos factores se han intentado dar al problema. La *perspectiva Bayesiana* se acomoda perfectamente a la hora de aportar una solución probabilística, teniendo en cuenta los conocimientos "a priori" de la población. Sin embargo, la modelización del comportamiento de un intruso no es un problema trivial y requiere modelos más complicados.(*)

Los modelos matemáticos demasiado complejos, llevan a veces a soluciones poco prácticas o difíciles de implementar. La mayor parte de las técnicas que trataremos a continuación estarán basadas en los denominados *métodos heurísticos*, entendiendo como tales aquellos que, basados en modelos teóricos consistentes (se puede demostrar la existencia o no de una solución óptima), aportan algoritmos convergentes (llegan a la solución óptima, si ésta existe, en un número finito de pasos o iteraciones). La principal desventaja de estas técnicas reside en que, en muchos casos, no se puede asegurar su eficiencia en tiempo.

Riesgo de Divulgación y Pérdida de Información

Otros dos conceptos fundamentales y estrechamente relacionados dentro del universo de la protección de datos, son el *riesgo de divulgación* y la *pérdida de información*. El principal objetivo de las técnicas de protección de datos consiste en aportar la máxima seguridad con un mínimo de información perdida. Medidas cuantitativas y cualitativas de estos dos aspectos se darán de forma que se pueda comparar la efectividad de los distintos métodos y su repercusión en la calidad y utilidad de la información publicada.

(*) Duncan, G. and Lambert, D. (1986) "Disclosure Limited Data Dissemination" Journal of Bisiness and Economic Statistics,81,10-18 ; (1989) "The Risk of Disclosure for Microdata" Journal of Business and Economic Statistics,7,207-217

Clasificación de las Técnicas de Protección de Datos

Las técnicas de protección de datos deben englobar las tres fases del proceso de producción de datos estadísticos*: la recogida, el procesamiento o análisis y la difusión.* Sin embargo, las más extendidas y desarrolladas se aplican en la última fase del proceso por lo que se denominan *Técnicas de Control de la Divulgación Estadística*. Éstas se clasifican de una manera casi natural, según el formato en el que los datos son publicados o difundidos. Comúnmente, existen tres principales formas de difundir los datos estadísticos:

- Mediante *Ficheros* de registros individuales.

- Mediante *Tablas* de magnitud o frecuencias.

- Mediante consultas secuenciales en *Bases de Datos.*

A continuación describiremos brevemente las técnicas en uso para cada una de las modalidades de difusión. Algunas de ellas serán tratadas detalladamente en este cuaderno, por su interés y aplicación práctica.

Técnicas para la protección en Ficheros de Datos

Con ficheros de datos estadísticos nos referimos a *ficheros de registros individuales que contienen información identificativa sobre cada individuo encuestado o registrado de una población o muestra*. El riesgo de divulgación de información confidencial en un fichero de datos vendrá dado principalmente por el *riesgo de identificación*.

Atributos tan claramente identificativos como nombres o direcciones, no son normalmente incluidos en ficheros de uso público y externo, por considerarse de carácter confidencial. Sin embargo características tales como sexo, edad, estado civil, población etc., sí son dadas a conocer junto con otras variables de interés. De esta forma, es posible identificar a individuos o unidades de la muestra o de la población que son únicos con respecto a una determinada combinación de valores de dichas variables. Como consecuencia de dicha identificación, será posible conocer toda aquella información registrada sobre el individuo en el propio fichero o en otros ficheros externos disponibles, de la misma población o muestra.

Las técnicas de protección en ficheros de datos estarán encaminadas en su mayoría a proteger contra la identificación de unidades o individuos únicos o "raros" en la población y evitar la difusión de valores inusuales de variables que favorezcan dicha identificación.

Muchas veces no disponemos de los datos de toda la población, sino de una muestra representativa de la misma, de forma que individuos o unidades que son raros o únicos en la muestra no tienen por qué serlo en la población. La estimación de la *proporción de unidades o individuos únicos en la población* a partir de los que sí lo son en la muestra o en el fichero de datos, puede convertirse en una medida de la efectividad de los métodos aplicados, es decir, en una *medida cuantitativa del riesgo de divulgación.*

Podemos clasificar estos métodos en dos grandes grupos:

- **Métodos de Restricción**. Se basan en limitar la cantidad de información publicada mediante diferentes técnicas:

- *Recodificación*: Consiste en unificar categorías en variables cualitativas o agrupar magnitudes con valores extremos en variables cuantitativas (top-bottom coding)

- *Supresiones Locales*: Simplemente se suprimen para su publicación, determinados valores sensitivos de variables que pueden dar lugar a identificaciones (profesiones poco comunes, salarios inusualmente altos,...).

- **Métodos de Perturbación.** Tienen en común que modifican los valores de determinadas variables, permitiendo su publicación pero de forma que no se puedan conocer los valores exactos. Entre éstos destacan:

  - *Redondeo Aleatorio.* Los valores de determinadas variables son sustituidos por cantidades redondeadas.

  - *Aportación de ruido.* Consiste en la introducción de error (pequeñas cantidades seleccionadas aleatoriamente) en los valores de las variables.

  - *Sustitución.* Consiste en intercambiar valores dentro de una variable de forma que la información puede ser tratada estadísticamente (se mantiene la estructura de correlación), sin riesgo de identificar a un individuo con un determinado registro dentro del fichero de datos.

  - *Microagregación.* Se agrupan los valores de la variable de acuerdo a determinados criterios de ordenación y en cada registro el valor de dicha variable es sustituido por la media del grupo al que pertenezca. Detalles sobre el método y su aplicación van a ser desarrollados con posterioridad en este cuaderno.

Mención especial merecen las ***técnicas de encriptación***, a las que no hemos englobado en ninguno de los grupos anteriores por diferir, tanto en su implementación como en el objeto final de su aplicación. Estas técnicas pueden ser aplicadas en cualquiera de las etapas del proceso de datos *(recogida, análisis y difusión)* y proporcionan una solución eficiente al problema de la *delegación segura* de datos (permite que la información sea analizada o procesada por agentes externos sin que éstos conozcan los valores de las variables que se están utilizando durante el proceso). También permiten la transferencia de ficheros de unos sistemas de información a otros de forma segura, sin que éstos puedan ser interceptados y descifrados por algún intruso. Una técnica criptológica será detallada en este cuaderno como solución al problema de la delegación segura de datos.

Técnicas para la protección de Tablas

La publicación y difusión de datos mediante tablas, bien sean de magnitudes agregadas o de frecuencias, es un sistema enormemente extendido, por lo que precisa de sus propias técnicas de protección que doten de seguridad, a la vez que no limiten la capacidad de informar eficientemente al usuario.

El hecho de que los datos publicados en tablas estén agrupados o resumidos no limita el riesgo de divulgación de información confidencial. Valores pequeños de las celdas en tablas de frecuencias o contribuciones dominantes al valor de la celda en tablas de magnitud, pueden aportar información sensitiva sobre los individuos o unidades que contribuyen a la celda.

El riesgo de divulgación en celdas vendrá dado por las denominadas *medidas de sensitividad*, que determinarán si una celda es o no sensitiva. El total de celdas sensitivas en una tabla determinará el riesgo total de divulgación de la misma. A mayor nº de celdas sensitivas, más insegura será la tabla.

Muchos de los métodos anteriormente explicados para el caso de protección de ficheros, pueden ser aplicados o adaptados al caso de datos tabulares. A continuación se resumen algunos de los más importantes:

- ***El redondeo controlado***. Difiere del *redondeo aleatorio* en que debe ser aplicado a las celdas de la tabla manteniendo la aditividad entre filas y columnas, es decir, comprobando la consistencia con totales y subtotales.

- ***Recodificación***. Como en el caso anterior, consiste en la agrupación de categorías en variables cualitativas o cuantitativas, de forma que el nuevo valor para la celda sea lo suficientemente grande como para considerarse seguro.

- ***Granularidad***. Se trata más de un criterio de seguridad que de una técnica de protección en sí misma. Se basa en la distribución de las celdas unitarias (con valor 1 en tablas de frecuencias o con contribuciones únicas en tablas de magnitud) dentro de la tabla. Si éstas sobrepasan un determinado nivel o *ratio de granularidad*, la tabla es considerada insegura. Debido al paralelismo de esta técnica con la identificación de unidades o individuos que son únicos o raros en ficheros de datos, podremos utilizar alguna de las técnicas explicadas en ese apartado, para proteger tablas con altos índices de granularidad.

- ***Supresión de celdas***. En datos tabulares no es suficiente con suprimir localmente las celdas consideradas como sensitivas ya que, debido a la existencia de totales, subtotales y el resto de valores de las celdas, es posible en muchos casos recalcular el valor de una celda confidencial. Es por tanto necesario realizar una *supresión secundaria de celdas* que, aun no siendo sensitivas, ayudan a calcular el valor de aquellas que sí lo son. Encontrar un patrón de supresión óptimo (es decir que aporte seguridad a la tabla con el mínimo número de supresiones) no es un problema trivial y precisa de técnicas de programación lineal que pueden llegar a complicarse para dimensiones grandes de la tabla (del orden de 4 en adelante).

Estas dos últimas técnicas serán desarrolladas con detalle a lo largo de este cuaderno, así como la explicación de diversos criterios de sensitividad y medidas de la pérdida de información que serán aplicados según el caso.

Técnicas para la protección en Bases de Datos

Una Base de Datos aporta un entorno rico en información en el que no sólo se encuentran almacenados datos estadísticos, sino los nexos y relaciones existentes entre ellos. El fácil acceso a esta información por parte del usuario mediante *sistemas de consultas secuenciales*, supone un alto riesgo de divulgación de información sensitiva.

Muchas bases de datos cimientan su seguridad, exclusivamente, en sistemas de acceso basados en claves o en permisos especiales para determinados usuarios. Sin embargo, existen múltiples formas de llegar a inferir datos sensitivos, bien sea por información obtenida en consultas previas o en otros entornos de bases de datos, o bien por los conocimientos previos del usuario sobre la población. La resolución de este problema inferencial en bases de datos, requiere de técnicas capaces de controlar, en

tiempo real, las consultas realizadas por un usuario y determinar si éstas pueden ser o no contestadas con seguridad.

En el presente cuaderno trataremos principalmente dos técnicas para la seguridad en bases de datos que tienen en cuenta los aspectos mencionados:

- ***Confidencialidad via Camuflaje (C.V.C.)***. Consiste básicamente en "esconder" la información confidencial entre dos valores límite que se dan como respuesta a la consulta realizada.

- ***Auditorías de Bases de Datos***. Controlan las consultas realizadas por un usuario y determina si la nueva consulta junto con la información aportada por las demás supone un riesgo de divulgación de datos sensitivos o confidenciales.

Procedimientos Informáticos

Además de analizar las diferentes técnicas de protección existentes, se realizará un análisis del funcionamiento del paquete informático ARGUS. Este software ha sido especialmente diseñado para la producción de datos seguros y consta de dos módulos: μ-Argus, para la protección de ficheros y τ-Argus para la protección de tablas. Ambos módulos han sido desarrollados en su mayoría por Statistics Netherlands y forman parte del proyecto europeo para el Control de la Divulgación Estadística.

Se trata por otro lado, del primer producto de estas características que proporciona un entorno amigable (funciona bajo Windows 95' y 98') y fácil de usar, a la vez que aplica de forma eficiente las técnicas de protección más en uso en estos momentos.

A lo largo del cuaderno y dentro de la sección que corresponda, se mostrarán sendos ejemplos de aplicación de ambos módulos junto con un breve desarrollo teórico de los modelos en los que se basa su funcionamiento.

Capítulo

2

Protección en Bases de Datos

Confidencialidad Via Camuflaje

Se trata de un método práctico propuesto por Gopal & Goes [12], orientado a garantizar la confidencialidad de datos numéricos en bases de datos de cualquier tamaño. Para su aplicación no es necesario conocer la distribución estadística que siguen dichos datos o realizar ninguna hipótesis inicial sobre la misma. Este método está encaminado a dar respuesta a cualquier tipo de consulta que el usuario externo de la base de datos quiera realizar, de forma correcta e ilimitada, sin comprometer información confidencial.

Aproximación al problema

El problema de la respuesta a consultas numéricas en bases de datos ya había sido tratado previamente mediante distintos métodos utilizados en protección de ficheros y adaptados para el caso concreto de consultas en bases de datos (métodos de Perturbación y de Restricción).

La Confidencialidad Via Camuflaje es una técnica que incorpora las ventajas de los métodos anteriores, eliminando sus mayores desventajas. Permite realizar un número ilimitado de consultas proporcionando respuestas correctas. Las respuestas son dadas en forma de un número más una garantía de manera que el usuario pueda determinar un intervalo dentro del cual se encuentra con seguridad la contestación exacta. La confidencialidad es mantenida "escondiendo" el vector de datos sensitivos en un conjunto infinito de vectores. Las ventajas de este método son notables:

- La técnica no depende de la distribución estadística que siguen los datos.
- Teóricamente, cualquier tipo de consulta puede ser contestada (respuesta ilimitada).
- Se puede aplicar en bases de datos de cualquier tamaño ya sean estáticas o dinámicas.
- El administrador de la base de datos puede controlar, basándose en los datos no confidenciales, el tipo de consultas que aporta los intervalos de respuesta más estrechos y por lo tanto más próximos a la respuesta exacta.

El Modelo

Consideramos la base de datos formada por n registros correspondientes a cada individuo o unidad de la población recogida en dicha base de datos. Los campos (características para cada individuo o unidad) de la base de datos se definen como confidenciales o no confidenciales. Para el análisis consideraremos la existencia de un único campo numérico confidencial. Así pues tendremos el siguiente vector de información confidencial:

$a = (a_1, a_2, \ldots\ldots\ldots a_n)$

El administrador de la base de datos, con libre acceso a los registros, puede determinar unos límites inferior (l_i), superior (u_i) o ambos para cada a_i, de forma que el campo se considerará protegido si con las sucesivas consultas que un potencial usuario puede realizar, éste no puede determinar que:

$$l_i < a_i \quad \text{ó} \quad u_i > a_i \quad \text{ó} \quad l_i < a_i < u_i$$

Cada consulta seleccionará un conjunto de registros $T \subseteq N = \{1,2,\ldots\ldots\ldots n\}$ que satisfacen unas determinadas condiciones asociadas a uno o más campos. Si dicha consulta no implica a campos confidenciales se denominará *consulta benigna.* Cada vez que se haga referencia a una consulta en el resto del análisis, consideraremos que ésta es *no benigna*, es decir, que implica algún campo confidencial.

La contestación a las consultas se dará en forma de respuesta puntual (*r*) además de una garantía (*g*), de forma que si denominamos como *e* a la respuesta exacta podremos decir siempre que $e \in I = [r^-, r^+] = [r-g, r+g]$.

Sería deseable obtener un *r* lo más cercano posible a *e* con *g* pequeño, siempre que el usuario no pueda determinar que $a_i \in (l_i, u_i)$. El objetivo final de esta técnica pretende otorgar respuestas lo más buenas y aproximadas posibles a las consultas asegurando la protección de los datos confidenciales.

La Técnica

Con el método CVC (Confidencialidad Via Camuflaje) la seguridad de los datos es mantenida incluyendo el vector confidencial **a** en un conjunto infinito de vectores P. Todas las respuestas a consultas pertenecen a este conjunto infinito, por lo que el usuario no puede saber más del vector **a** que su pertenencia a P. Además para cada individuo o unidad i, al menos uno de los vectores de P contiene un número que no excede el límite inferior l_i y otro que contiene un número que no es menor que el límite superior u_i de forma que el usuario nunca puede saber que $a_i \in (l_i, u_i)$.

Consideramos el conjunto de vectores:

$$P = \{P^1, P^2, \ldots\ldots\ldots, P^{k-1}, P^k\} \quad \text{con} \quad P^j = \{p_1^j, p_2^j, \ldots\ldots, p_n^j\} \quad \text{y} \quad P^k = a = (a_1, a_2, \ldots\ldots\ldots, a_n)$$

Sean $p_i^- = \min_{j=1,\ldots,k} p_i^j$ y $p_i^+ = \max_{j=1,\ldots k} p_i^j$ que satisfacen:

$$p_i^- \leq l_i \qquad \text{si } l_i \text{ está especificado} \qquad (1)$$

$$p_i^+ \geq u_i \qquad \text{si } u_i \text{ está especificado} \qquad (2)$$

El vector confidencial **a** está "escondido" en el conjunto infinito $\mathrm{P} = conv(P)$.

El índice k representa el nº de campos numéricos de la base de datos y suele ser pequeño (del orden de $3 \leq k \leq 6$) en la mayoría de los casos.

Asumiendo que la mayor parte de las consultas numéricas pueden ser representadas mediante funciones f(**x**) (univariantes o multivariantes) se puede expresar $[r^-, r^+]$ como:

$r^- = r^l = \min f(\mathbf{x})$, $\mathbf{x} \in P$

$r^+ = r^u = \max f(\mathbf{x})$, $\mathbf{x} \in P$

De forma equivalente :

$$r^l = \min f(l) = \min f(\sum_{j=1}^{k} l_j \cdot P^j), \quad \sum_{j=1}^{k} l_j = 1 \quad l_j \geq 0, \quad j = 1,.....,k \tag{3}$$

$$r^u = \max f(l) = \max f(\sum_{j=1}^{k} l_j \cdot P^j), \quad \sum_{j=1}^{k} l_j = 1 \quad l_j \geq 0, \quad j = 1,.....,k \tag{4}$$

Dónde $l = (l_1, l_2,....., l_k)$. Se sigue que $\mathbf{e} \in [r^l, r^u]$ ya que $\mathbf{a} \in P$. La protección está garantizada por (1) y (2).

Para determinadas consultas las expresiones (3) y (4) son sencillas de calcular (p.ej para el caso en el que f(l) es lineal), luego la solución vendrá dada por el intervalo $[r^l, r^u]$. Si las expresiones no son computables habrá que responder con $[r^-, r^+]$ dónde $r^- \leq r^l$ y $r^+ \geq r^u$ para asegurar la protección. La clave estará en calcular estos valores de forma que la garantía **g** sea lo más pequeña posible.

La eficiencia del método se basa en que el número de veces que se accede a un determinado registro no es mayor que si la respuesta fuera dada directamente por f(**a**) (es decir, en función de los datos confidenciales).

Auditorías de Bases de Datos

En entornos interactivos de bases de datos es necesario mantener un cierto control sobre las *consultas* que un usuario puede realizar y si éstas implican, ya no sólo información sensitiva, sino valores que junto con otros de consultas anteriormente realizadas pueden aportar "pistas" que ayuden a aproximar muy estrechamente información confidencial. La auditoría de bases de datos es un método eficiente para evitar la difusión de información sensitiva y la inferencia de datos confidenciales y además utiliza un modelo matemático donde el número de variables nunca es mayor (en ocasiones mucho menor) que el tamaño de la base de datos [20].

Conceptos Previos

Antes de entrar de lleno en la implementación de una auditoría para bases de datos, es necesario familiarizarse con los términos que van a ser utilizados posteriormente. Para una determinada *consulta* hablaremos de:

- **Fórmula Lógica**

Denominada también *categórica o característica*, hace referencia a la construida uniendo mediante operadores booleanos ($\wedge,\vee,\neg$) y de condición, valores de atributos (campos de la base de datos), de forma que selecciona a un determinado conjunto de registros de la base de datos, denominado ***conjunto-consulta***.

- **Función de Agregación**

Es aquella que tomando como entrada un atributo (campo) numérico devuelve un *valor* que será el resultado de calcular la operación de que se trate (suma, media, frecuencia, máximo, mínimo,...) en el *conjunto-consulta.*

(Ver Ejemplo)

Criterios de Sensitividad

Cuando un sistema de bases de datos detecta que el valor de una determinada consulta puede ser sensitivo (es decir, puede aproximar el valor de un atributo confidencial), denegará la respuesta a la misma en base a unas reglas de sensitividad que definimos a continuación y que variarán de un sistema a otro según sea el tipo de datos y consultas:

- **Regla del Valor-Límite (para consultas contador)**

La sensitividad viene determinada por un número entero positivo n, de forma que si la consulta implica a n o menos registros entonces se considera sensitiva. Este valor es elegido de forma que la probabilidad de identificar algún registro como miembro de un conjunto arbitrario de n+1 o más registros es aceptablemente pequeña con respecto a un criterio dado.

- **Regla de Dominancia (N, K) (para consultas suma)**

Considera que una consulta q es sensitiva si N o menos registros del conjunto-consulta constituyen más del K% de la suma total.

Inferencia de un conjunto de consultas

Aun cuando el valor de la consulta realizada no sea sensitivo, contestar a cualquier consulta limitando la protección únicamente al criterio de sensitividad, deja al azar la posible difusión de datos confidenciales que pueden ser inferidos por el usuario gracias a información aportada por anteriores consultas. Así pues, hablaremos de conjuntos de consultas "seguros" y de métodos de inferencia que nos ayudarán a decidir si contestar a una nueva consulta compromete de alguna manera datos confidenciales.

Definición 1. Dada una secuencia de consultas respondidas $\{q_1,q_2,....q_n\}$ considerada segura, una nueva consulta q_{n+1} decidirá si q_{n+1} puede ser contestada con seguridad.

Notación:

Sea Q un conjunto de **consultas suma** y S_q los registros seleccionados por la consulta $q \in Q$ (ó *conjunto-consulta* de q). Llamaremos R a la unión de todos los S_q de Q y t_q al valor resultante de la consulta q.

Creamos una partición única π de R formada por clases disjuntas de manera que cada S_q no vacío es la unión de una o más clases de π. Los valores t_q podrán ser resumidos en un sistema de restricciones lineales cuyas variables corresponden 1-1 a las clases de π. Estas clases serán exactamente los valores de las $2^{|Q|}$ -1 expresiones no vacías del tipo:

$$\bigcap_{q\in Q} s_q \quad \textit{dónde } s_q \textit{ es } S_q \textit{ ó } \quad \overline{S}_q = R - S_q \textit{ con la exclusión de } \bigcap_{q\in Q} \overline{S}_q$$

Denominaremos *diagrama consulta* H de Q al hipergrafo cuyos vértices son cada elemento de Q y aristas dadas por el conjunto:

$$E = \{e \subseteq Q \mid (\bigcap_{q\in e} S_q) \cap (\bigcap_{q\notin e} \overline{S}_q) \textit{ es una clase de } p\}$$

Denotaremos x(*e*) a la variable correspondiente a cada arista *e* de H y formaremos el sistema de restricciones **Mx=t**, dónde M es la matriz vértice-arista de H.

Sea A un subconjunto de E. Consideramos la siguiente expresión suma, de la cual A será el *soporte* : $x(A) = \sum_{e\in A} x(e)$

Diremos que x(A) es un *invariante* del sistema si x(A) es constante, es decir, que para cada dos soluciones $\mathbf{x_1}$ y $\mathbf{x_2}$ del sistema se tiene que:

$$\sum_{e\in A} x_1(e) = \sum_{e\in A} x_2(e)$$

Definición 2. Un conjunto de aristas de H se dice que es un *conjunto invariante* si es el soporte de un invariante del sistema.

Definición 3. Dado un conjunto de registros S, sea X el dato obtenido al evaluar la función de agregación en todos los registros de S. Decimos que X *se puede inferir* de Q si S=∅ ó existe un subconjunto no vacío A, del conjunto de aristas del diagrama-consulta de Q, tal que:

$(i)\quad S = \bigcup_{e \in A}[(\bigcap_{q \in e} S_q) \cap (\bigcap_{q \notin e} \overline{S}_q)$ (Pr*opiedad de cobertura*)

(ii) A es un conjunto invariante de H (Pr*opiedad de invarianza*)

Así pues un conjunto de consultas se considerará seguro si la información (X) que se puede inferir del sistema de restricciones lineales construido a partir de él, no es sensitiva.

Un algoritmo basado en el cumplimiento de las dos propiedades anteriores nos permitirá decidir si podemos contestar con seguridad a una determinada consulta, sin que esto implique la posible inferencia de datos sensitivos.

Test de respuesta

Este test ayudará a decidir para un conjunto de consultas contestadas $\{q_1,q_2,....q_n\}$ que se considera seguro, si la respuesta a una nueva consulta q_{n+1} se puede realizar de forma segura y además determinará también si el nuevo conjunto $\{q_1,q_2,....q_n, q_{n+1}\}$ es seguro a su vez.

El test tomará como entrada un subconjunto Q de $\{q_1,q_2,....q_n\}$, al que denominaremos *base,* de forma que cualquier valor de $q_i \notin Q$ puede ser inferido del subconjunto $Q \cap \{q_1,q_2,....q_{i-1}\}$ y la información inferida no es sensitiva. Si para un q_{n+1} no sensitivo, se sigue cumpliendo esta propiedad, Q no variará y se podrá responder q_{n+1}. Además $\{q_1,q_2,....q_n,q_{n+1}\}$ será un conjunto de consultas seguro. Si no es así, es posible que exista un subconjunto de aristas de E (aristas sensitivas) que junto con el valor de q_{n+1} difunde información sensitiva. Para detectar esta circunstancia se creará una nueva base Q' para $\{q_1,q_2,....q_n, q_{n+1}\}$ y se realizará un test de seguridad, consistente en aplicar el test de sensitividad a cada arista sensitiva. Si el resultado de esta prueba es negativo ($\{q_1,q_2,....q_n, q_{n+1}\}$ no es seguro) no se contestará q_{n+1} , de lo contrario se dará a conocer q_{n+1} y Q' será la nueva base para posteriores consultas.

A continuación se resumen los pasos seguidos por el algoritmo:

Entradas: q_{n+1} ;H=(Q,E) es el diagrama-consulta de Q con E=conjunto de aristas de Q.

Paso 1. *Test de sensitividad* para q_{n+1}

Si q_{n+1} es sensitivo se deniega la respuesta y **Fin.**

Si_no Paso 2.

Paso 2. *Test de inferencia.*

Si H cumple (*prop. de cobertura* Ù *prop. de invarianza)* entonces q_{n+1} se puede inferir de Q. Se dará a conocer el valor de q_{n+1} y la base de $\{q_1,q_2,....q_n, q_{n+1}\}$ seguirá siendo Q.**Fin.**

Si_no Paso 3.

Paso 3. Se construye un nuevo diagrama H' para $Q \cup \{q_{n+1}\}$.

Paso 4. *Test de seguridad* para H'

Si el test es negativo entonces se deniega la respuesta y **Fin.**

Si_no se publica q_{n+1} y $Q' = Q \cup \{q_{n+1}\}$ será la nueva base para posteriores consultas.

Aunque se aporta este algoritmo, aún es necesario optimizar el proceso de detección de aristas sensitivas (Test de seguridad) [(*)] de forma que se pueda generalizar para casos diferentes al de datos reales.

Ejemplo

Vamos a identificar en el siguiente ejemplo propuesto en [20] algunos de los conceptos descritos:

Consideramos la base de datos EMPRESA que contiene 9 registros correspondientes a 9 hipotéticos trabajadores sobre los que se ha almacenado los atributos ID(clave de identificación del trabajador), DEP (departamento en el que trabaja) y SALARIO (sueldo del trabajador).

Registro	ID	DEP	SALARIO
1	id1	Dirección	85
2	id2	Dirección	3
3	id3	Dirección	2
4	id4	Administración	4
5	id5	Administración	3
6	id6	Administración	3
7	id7	Servicios	4
8	id8	Servicios	3
9	id9	Servicios	3

Se realizan las siguientes consultas (en SQL) sobre la base de datos EMPRESA:

***q_1**: select SUM(SALARIO) from EMPRESA where DEP=Administración*

Valor de $q_1 = t_1 = 10$

***q_2**:select SUM(SALARIO) from EMPRESA where DEP=Servicios*

Valor de $q_2 = t_2 = 10$

***q_3**:select SUM(SALARIO) from EMPRESA*

Valor de $q_3 = t_3 = 110$

***q_4**:select SUM(SALARIO) from EMPRESA where DEP=Dirección*

Valor de $q_4 = t_4 = 90$

[(*)] Para más información sobre el Test de seguridad ver Malvestuto,F.M., Moscarini, M. (1990) "Query evaluability in statistical databases" IEEE Transactions on knowledge and data engineering 2, 425-430.

Si aplicamos la regla dominancia (N, k)=(2, 85%) diremos que la consulta q_4 es sensitiva ya que existe una contribución (la del registro 1) que representa más del 85% del valor de la consulta. El resto de consultas no son sensitivas con respecto a este criterio.

A continuación identificamos las fórmulas lógicas y de agregación y el conjunto-consulta para cada q_i no sensitiva:

- *Función de agregación*: SUM(SALARY) en todos los casos.
- *Fórmulas lógicas o categóricas* *Conjunto-consulta*

  q_1: DEP=Administración $S_1=\{4,5,6\}$

  q_2: DEP=Servicios $S_2=\{7,8,9\}$

  q_3: Verdadero(True) $S_3=\{1,2,3,.....,9\}$

Determinamos los conjuntos Q y R, la partición π y el conjunto de aristas E:

$Q=\{q_1,q_2,q_3\}$ $R=\bigcup S_i = S_3$

Las clases para la partición π vienen dadas por las siguientes expresiones (las únicas no vacías de las 2^3-1 posibles):

$$S_1 \cap \overline{S}_2 \cap S_3 = S_1 = \{4,5,6\}$$

$$\overline{S}_1 \cap S_2 \cap S_3 = S_2 = \{7,8,9\}$$

$$\overline{S}_1 \cap \overline{S}_2 \cap S_3 = S_3 - (S_1 \cup S_2) = \{1,2,3\}$$

El conjunto de aristas $E=\{e_1,e_2,e_3\}=\{ (q_1,q_3), (q_2,q_3), (q_3) \}$ cumpliéndose que:

$$Para\ e_1 : (S_1 \cap S_3) \cap \overline{S}_2 = \{4,5,6\}\ es\ una\ clase\ de\ p$$

$$Para\ e_2 : (S_2 \cap S_3) \cap \overline{S}_1 = \{7,8,9\}\ es\ una\ clase\ de\ p$$

$$Para\ e_1 : S_3 \cap (\overline{S}_1 \cap \overline{S}_2) = \{1,2,3\}\ es\ una\ clase\ de\ p$$

Sea x_i la variable asociada a cada una de las clases de π, el sistema de ecuaciones lineales Mx=t que resume la información aportada por las respuestas a las consultas q_1,q_2 y q_3 es el siguiente:

$$\begin{cases} x_1 = 10 \\ x_2 = 10 \\ x_1 + x_2 + x_3 = 110 \end{cases} \quad \text{con } x_i \geq 0$$

El valor de x_3 (que será la suma de los salarios para el departamento de dirección) queda determinado de forma única (x_3=90) por el sistema. De este modo y por tratarse de un valor sensitivo (según la regla de dominancia utilizada) podrá aproximarse muy estrechamente el valor del salario del empleado id1.

Así pues podemos considerar que Q no es un conjunto de consultas seguro, ya que información sensitiva puede ser inferida de las respuestas dadas. Sería necesario aplicar el test de respuesta antes de contestar q_3, que no siendo un valor sensitivo, sí aporta información, con la ayuda de valores de anteriores consultas, sobre datos confidenciales de la base de datos.

Capítulo

3

Protección en Ficheros de Datos

Microagregación

La microagregación es una técnica de control de la divulgación estadística en ficheros de datos que básicamente consiste en agrupar los registros individuales en pequeños estratos antes de su publicación o difusión en el mercado. Hasta ahora, en la práctica, se viene aplicando la *microagregación* de tamaño fijo, es decir, todos los pequeños grupos que se forman tienen el mismo tamaño k. En este apartado, trataremos algunas de estas técnicas, pero también abordaremos el análisis de métodos dónde el tamaño de los estratos toma valores variables ($\geq$ k) dependiendo de la distribución inicial de los datos [4].

Aproximación al problema

La microagregación se basa en la existencia de ciertas reglas de confidencialidad que permiten la publicación y difusión de datos, siempre que éstos estén agrupados de forma que ningún dato individual destaque sobre los demás. La estricta aplicación de estas reglas lleva a reemplazar valores individuales de variables por valores calculados por pequeños estratos (medias para cada grupo).

Para obtener estos estratos, consideramos un conjunto de vectores (uno por cada registro del conjunto de datos) y los combinamos para formar grupos de, por lo menos, tamaño k. Dentro de cada grupo, se calcula el valor de la media para cada una de las variables numéricas que reemplazará al valor original.

Notación:

Consideramos un conjunto de datos con **p** variables continuas y **n** vectores de datos (registros) resultantes de observar las p variables en n individuos. Con estos vectores formamos **g** grupos con $\mathbf{n}_i$ individuos en cada grupo i-ésimo ($n_i \geq k$ y $n = \Sigma\, n_i$).

Denotamos:

$\mathbf{X}'=(x_1,x_2,.....,x_p)$ vector de datos dónde x_i son valores de variables.

$\mathbf{X}_{ij}$ j-ésimo vector del grupo i $j=1,.....,n_i$

$\grave{\mathbf{X}}_i$ vector de medias del grupo i $i=1,.....,g$

$\grave{\mathbf{X}}$ vector de medias del conjunto de datos

La formación de grupos se basa en criterios de máxima similitud. Este es el principal problema de la k-partición, ya que es necesaria una medida de similitud entre vectores. Cada vector individual puede ser visto como un punto y todo el conjunto de datos como un conjunto de puntos multidimensionales (la dimensión dependerá del número de

variables medidas dentro de cada vector). De esta forma, la semejanza entre vectores puede ser medida utilizando una *distancia.*

Suma de Cuadrados dentro de los grupos

$$SSE = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x_{ij}} - \bar{\mathbf{x}}_\mathbf{i})'(\mathbf{x_{ij}} - \bar{\mathbf{x}}_\mathbf{i})$$

Suma de cuadrados entre grupos

$$SSA = \sum_{i=1}^{g} n_i(\bar{\mathbf{x}}_\mathbf{i} - \bar{\mathbf{x}})'(\bar{\mathbf{x}}_\mathbf{i} - \bar{\mathbf{x}})$$

Suma de cuadrados Total

$$SST = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x_{ij}} - \bar{\mathbf{x}})'(\mathbf{x_{ij}} - \bar{\mathbf{x}})$$

La partición óptima será aquella que minimice la suma de cuadrados dentro de los grupos (SSE) o de forma equivalente la que maximice la suma de cuadrados entre grupos (SSA). Estas sumas de cuadrados pueden utilizarse como *medida de la pérdida de información*:

$$L = \frac{SSE}{SST} \qquad (1)$$

El valor de L se encontrará entre 0 y 1. Cuanto más cercano sea a 0 menor será la pérdida de información debida a la microagregación. Los métodos heurísticos presentados a continuación, son una alternativa práctica que trata de minimizar la pérdida de información mediante el conocimiento de la variabilidad de los datos (SST *y* SSE).

Métodos Unidimensionales con Tamaño Fijo

Estos métodos heurísticos ordenan de forma ascendente o descendente los vectores de datos de acuerdo con un criterio unidimensional determinado. Una vez ordenados, se forman grupos de *tamaño fijo k* con vectores sucesivos. Si el número total de vectores n no es múltiplo de k, entonces el último grupo contendrá más de k vectores. Dentro de cada grupo se reemplaza el valor de cada variable por su media dentro del grupo.

- **Métodos del Eje-Único**

Resultan muy efectivos cuando las variables están fuertemente correladas. La ordenación (ascendente o descendente) se puede realizar siguiendo varios criterios:

- *Ordenación por componentes principales.* Se ordenan los vectores por la 1ª componente principal, siendo ésta una variable transformada de forma que está altamente correlada con el resto de variables originales.

- *Ordenación por una variable particular*. Dicha variable debe reflejar de alguna forma el tamaño del vector de datos.

- *Ordenación por la suma de los z-scores*. Todas las variables son estandarizadas y para cada vector se calculan las sumas de dichos valores estándar, ordenando por esta característica.

- **Método de Ordenación Individual**

Cada variable se considera de forma independiente de las demás. Los vectores son ordenados por la 1ª variable, formándose grupos de k vectores sucesivos y reemplazando los valores de cada variable por su media dentro de cada grupo. Un procedimiento similar se lleva a cabo con el resto de variables (ordenando por la 2ª, 3ª,...y así sucesivamente).

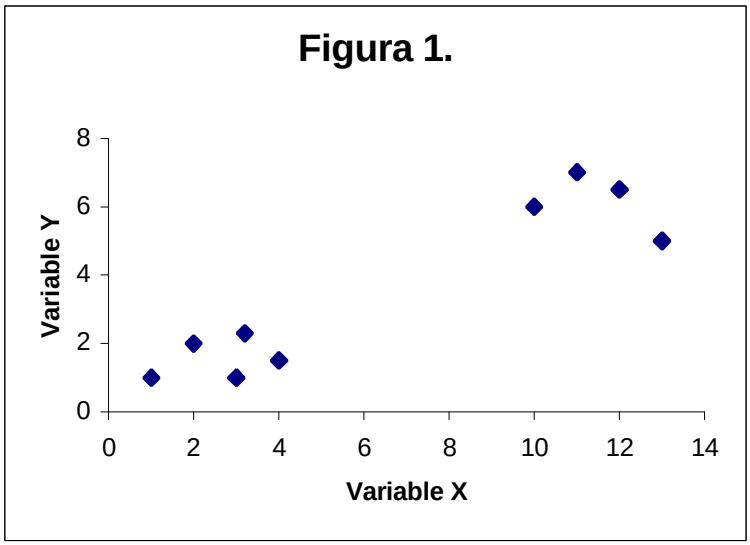
Este método preserva gran cantidad de información pero existe un mayor riesgo de difusión de datos que pueden ser confidenciales. El usuario puede llegar a deducir que el verdadero valor de una variable en el grupo i se encuentra entre la media del grupo i-1 y la del i+1. Si dichas medias tienen valores próximos, el intervalo que se conoce puede ser demasiado estrecho.

Otra desventaja de éste método consiste en que los n vectores no son agrupados en un único vector de medias básico, ya que la microagregación se realiza para cada variable cada vez por lo que se obtiene una partición diferente (y por lo tanto un vector de medias distinto) para cada variable del conjunto de datos.

Métodos con Tamaño Variable

Hasta ahora las técnicas de microagregación más aplicadas suelen utilizar grupos de vectores de tamaño fijo. Sin embargo, la posibilidad de variar el tamaño de un grupo a otro de manera que dicha agrupación se adaptara mejor a la distribución inicial de los datos, tendería a perder menos información (conservando siempre la condición de ser al menos de tamaño k). La siguiente figura ilustra las ventajas de la microagregación con grupos de tamaño variable:

Figura 1.



El gráfico muestra dos variables y nueve datos. Si utilizáramos microagregación de tamaño fijo k=3, obtendríamos una partición de datos en tres grupos, lo que parece poco natural para la distribución dada. Si por el contrario utilizamos microagregación de tamaño variable ≥ k, los cinco datos a la izquierda del gráfico podrían ser incluidos en un mismo grupo y los cuatro restantes en otro. De esta forma la agrupación de vectores resulta más acorde con la distribución de los datos y la pérdida de información es menor.

Dos métodos de microagregación con tamaños variables son presentados a continuación:

- **Microagregación Genética**

Debe su nombre a que representa las particiones como cadenas binarias (al igual que cromosomas) y combina una búsqueda directa y aleatoria para encontrar la agrupación óptima. Se trata de un método difícil de adaptar al caso multivariante ya que un espacio multidimensional está ordenado sólo parcialmente por lo que la representación binaria no resulta la más adecuada.

- **Método de Ward(*)**

Se trata de un método bastante efectivo y proporciona un algoritmo recursivo que optimiza la solución en cada paso. En cada iteración se unen dos vectores (ó grupos de vectores), de forma que el incremento de la suma de cuadrados (SSE) debida a su unión, sea mínimo.

El método de Ward surge originariamente, como un método de agrupación jerárquica de vectores que optimiza SSE, sin considerar ningún tamaño mínimo k para los grupos. Por ello, ha sido adaptado para trabajar con microagregación introduciendo ligeras modificaciones en lo que denominamos el *Algoritmo Modificado de Ward*.

Algoritmo Modificado de Ward

El algoritmo modificado de Ward proporciona un método de microagregación para datos numéricos o cualitativos (si ha sido definido previamente un concepto de distancia). A continuación serán explicados brevemente(**) los pasos que sigue para *el caso univariante*, de forma que se pueda entender su funcionamiento y la mejora que supone con respecto a otros métodos unidimensionales. El salto al *caso multivariante* sólo supondrá algún ligero cambio en el algoritmo que también será detallado en esta sección.

Antes de pasar a detallar el algoritmo, son necesarias ciertas definiciones previas:

Definición 1. *Para un conjunto de datos dado, una k-partición P es cualquier partición del conjunto de datos de forma que cada grupo de P contenga por lo menos k elementos.*

(*) Ward,J.H.(1963) "Hierarchical grouping to optimize an objective function" *Journal of the American Statistical Association, 58, 236-244*

(**) Para una explicación más amplia y detallada ver Domingo,J., Mateo, J.M. (1997) "Practical Data-Oriented Microaggregation for Statistical Disclosure Control" *Journal of Classification*

Definición 2. *Para un conjunto de datos dado, la k-partición P se dice que es* ***"mejor que"*** *la k-partición P' si cada grupo de P está contenido en algún grupo de P'.*

Definición 3. *Para un conjunto de datos dado, la k-partición P se dice que es* ***mínima*** *con respecto a la relación "mejor que" si no existe una k-partición P'¹P tal que P' es mejor que P.*

Proposición 4. *Para un conjunto de datos dado, la k-partición P se dice que es* ***mínima*** *con respecto a la relación "mejor que" si y sólo si está formada por grupos con tamaños ³ k y < 2k.*

Corolario 5. *Existe una solución óptima al problema de la k-partición, si ésta es mínima con respecto a la relación "mejor que".*

Algoritmo. Caso univariante

Paso 1. *Partiendo de un conjunto de datos* ***ordenado*** *se forman dos grupos; uno con los k primeros elementos (los más pequeños) y otro con los k últimos elementos (los más grandes).*

Paso 2. *Aplicar el método de Ward hasta que todos los elementos del conjunto de datos pertenezcan a un grupo de tamaño ³ k. En este paso, no debe unirse dos grupos si ambos tienen un tamaño ³ k.*

Paso 3. *Para cada grupo que en la partición resultante tenga más de 2k elementos, aplicar este algoritmo recursivamente.*

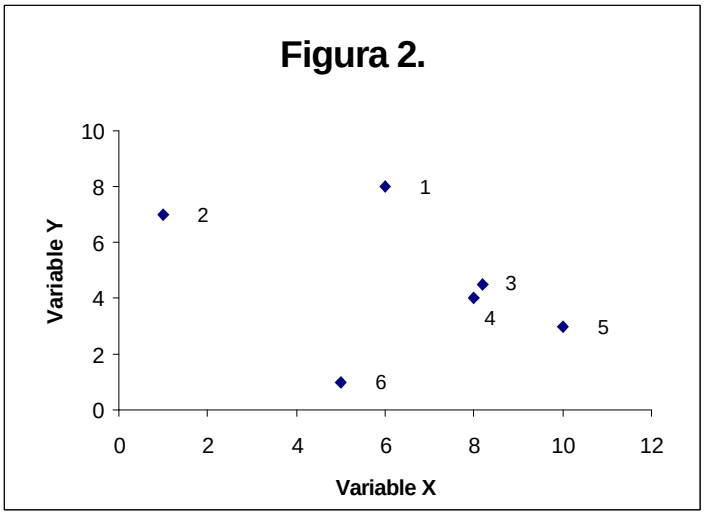
Se puede demostrar que el algoritmo termina en un número finito de pasos y obtiene la partición mínima para el k dado. Aún cuando no mantiene la optimalidad en cada paso (como ocurre en el método de Ward), esta técnica mejora los resultados obtenidos por los métodos unidimensionales en lo que se refiere a pérdida de información.

Algoritmo .Caso multivariante

Sólo el paso 1 del algoritmo ha de ser adaptado para el caso multivariante. Es necesario definir claramente lo que entendemos por los k "primeros" elementos y los k "últimos" elementos , ya que en este caso lo que agrupamos son vectores multidimensionales.

Una solución consiste en utilizar los métodos unidimensionales (Eje-Único, Ordenación Individual...) anteriormente explicados, no como técnicas de microagregación en sí mismas sino como criterios de ordenación de vectores en el paso 1 del algoritmo modificado de Ward .

No obstante, existe un método de ordenación adicional para el caso multivariante denominado *Criterio de Máxima Distancia.* Consiste fundamentalmente en seleccionar para cada uno de los vectores "extremo ", los k-1 vectores más próximos según la matriz de distancias. De esta forma se obtiene el grupo de los k "primeros " vectores y el de los k "últimos" vectores. La agrupación resultante según este criterio de ordenación será distinta dependiendo de cual de los vectores "extremo" elijamos como 1er o último vector. La siguiente figura nos aclara este último punto:



Si consideramos los seis vectores bidimensionales de la figura 2 y tomamos k=3, se observa que los vectores más distantes (vectores "extremos") son los etiquetados como 2 y 5. Si comenzamos la ordenación por el vector 2, el vector más cercano a él es 1; y a su vez el más cercano al grupo (1,2) es el vector 3, luego obtendremos los grupos (1,2,3) y (4,5,6). Sin embargo si comenzáramos la ordenación por el vector 5 los grupos obtenidos serían claramente (3,4,5) y (1,2,6).

No obstante, las diferencias en términos de pérdida de información por elegir uno u otro vector como 1º o último no son muy grandes.

Una de las principales desventajas de este método, reside en la necesidad de almacenar una matriz que contenga la distancia para cada par de datos. Si tenemos un conjunto de n vectores de datos, la capacidad de almacenamiento por tratarse de una matriz simétrica y con ceros en la diagonal, ha de ser del orden de (n-1)n/2.

A modo de conclusión, resaltar que el Algoritmo Modificado de Ward mejora notablemente los resultados de los anteriores métodos heurísticos de microagregación, en lo que a pérdida de información se refiere. La línea de investigación futura irá por el camino de mejorar los resultados computacionales, tanto en tiempo como en capacidad de almacenamiento.

Criptografía

En muchas ocasiones los datos estadísticos necesitan ser procesados fuera de su lugar de origen o enviados a agencias estadísticas externas para su utilización con distintos fines. Si el propietario de los datos ha adquirido ciertos compromisos de confidencialidad, no puede delegar dicha información sin que antes sea tratada de forma que su procesamiento no implique la difusión de datos confidenciales.

Existen soluciones legales a este problema en forma de acuerdos de no-difusión en los que el firmante se compromete a no publicar ningún dato de carácter confidencial. Sin embargo, fuera del marco legal el propietario debe confiar en la honestidad de la agencia externa que procesará los datos, siempre que no exista algún medio técnico que controle el adecuado uso de los mismos.

En este apartado se aporta una solución criptográfica al problema de la *delegación segura de datos* [6], que permite controlar y limitar el tipo de operaciones que se pueden llevar a cabo con la información delegada.

Delegación segura de datos

Dos condiciones fundamentales deben darse en una delegación de datos segura:

- **El secreto de los datos**

El propietario debe asegurarse de que los datos confidenciales permanecen secretos. Una forma de procesar datos sin necesidad de actuar sobre la información original consiste en operar con datos previamente *encriptados.* Para ello el propietario de la información debe utilizar técnicas de encriptación denominadas *Homomorfismos Privados*, que permiten al agente externo realizar operaciones aritméticas básicas (sumas, restas, multiplicaciones, divisiones...), incluso tests de igualdad, sobre datos encriptados sin exponer la confidencialidad de los mismos.

- **Verificación del procesamiento**

El propietario ha de verificar que el uso que se le ha dado a los datos durante su procesamiento ha sido el correcto.

Un procedimiento para llevar a cabo la comprobación del procesamiento de datos, sin tener que repetir todo el proceso de cálculo realizado por el agente externo, consiste en realizar una verificación probabilística denominada *test de paridad.* La idea consiste en transformar operaciones aritméticas en expresiones booleanas obteniendo para cada expresión su *paridad* (es decir, calcular si es par o impar). Si denominamos Z(x) a la función de paridad, las expresiones de paridad de a+b y a∗b pueden ser calculadas como:

$$Z(a+b)= Z(a) \oplus Z(b)=Z(a) \textbf{ OR } Z(b)=Z(o)$$

$$Z(a*b)= Z(a) \cdot Z(b) = Z(a) \textbf{ AND } Z(b)=Z(y)$$

Si para cualquier dato de entrada d, tenemos que su distribución de paridad estimada es $(1-p_d,p_d)$ con p_d=probabilidad de que d sea impar, la distribución de paridad del resultado para las operaciones suma y producto se calcularán como:

$$Z(o) \text{ se distribuye como } (1-p_o,p_o) \quad \text{con } p_o=p_a*(1-p_b) + (1-p_a)*p_b$$

$$Z(y) \text{ se distribuye como } (1-p_y, p_y) \quad \text{con } p_y = p_a * p_b$$

De este modo se puede calcular la paridad del resultado de cualquier expresión booleana (con términos AND y OR). Por ello se requiere del agente externo, que proporcione junto con el resultado encriptado, la expresión utilizada en el proceso de los datos *(expresión requerida)*. El propietario desencripta el resultado *(expresión computada)* y calcula su paridad. Por otro lado, calcula la paridad resultante de sustituir en la expresión requerida los datos originales. Si el procesamiento de los datos ha sido el correcto ambas paridades deben coincidir, si difieren puede haber sido detectado un fraude.

Como ya hemos visto, las distribuciones de paridad de los resultados de las expresiones requerida y computada son respectivamente (1-p,p) y (1-p',p') luego la probabilidad de detectar un fraude viene dada por:

$$P(\text{detectar fraude}) = P(\text{expr. requerida} = \text{impar}) \cdot P(\text{expr. computada} = \text{par}) +$$

$$+ \; P(\text{expr. requerida} = \text{par}) \cdot P(\text{expr. computada} = \text{impar})$$

Lema. *Si la expresión requerida y la expresión computada difieren, la probabilidad de que el propietario de los datos detecte un fraude viene dada por la expresión:*

$$P(\text{detectar}) = p(1-p') + p'(1-p) \quad \text{con } p = \text{prob. de que la expresión requerida sea impar} \quad (1)$$

$$y \;\; p' = \text{prob. de que la expresión computada sea impar}$$

Teorema 1. *Si los datos de entrada se distribuyen de una forma aleatoria, la probabilidad de detectar un fraude mediante el test de paridad, está acotada superiormente por 1/2.*

Dem/ Si los datos de entrada son aleatorios su distribución de paridad es (1/2,1/2). La operación suma no introduce sesgo en la paridad, pero sí lo hace la operación producto (sólo en el caso impar $\times$ impar = impar), luego $p \leq 1/2$ y $p' \leq 1/2$ en (1), por lo tanto $P(\text{detectar}) \leq 1/2$. Si las expresiones requeridas y computadas sólo contienen sumas entonces $p = p' = P(\text{detectar}) = 1/2$.

Teorema 2. *Si los datos de entrada se distribuyen de una forma aleatoria, la probabilidad de detectar un fraude mediante el test de paridad, está acotada inferiormente por max(p,p'). Si las expresiones siempre pares están prohibidas por el propietario de los datos, entonces la probabilidad de detectar un fraude es siempre > 0.*

Una forma muy sencilla de detectar si existe alguna expresión siempre par (que se suponen prohibidas), consiste en igualar la fórmula de paridad del resultado a 0. Para facilitar esta operación, se solicita al agente externo que devuelva el resultado en forma de suma de productos, de forma que para cada producto (término AND) en la expresión booleana baste comprobar que existe el complementario que lo anula.

Se puede concluir, que el propietario de los datos puede asegurar que la probabilidad de detección de un fraude va a ser mayor que cero, sin embargo no podrá determinar la cota inferior mencionada en el Teorema 2 ya que desconoce p' (no conoce la expresión realmente utilizada por el agente externo, si éste ha cometido un fraude). Una forma de incrementar la probabilidad de detección de fraude, consiste en realizar el test de paridad con resultados intermedios, cuantas más expresiones intermedias se verifiquen mayor será la probabilidad de detección de fraude, aun cuando esto suponga un incremento del coste en tiempo por parte del propietario de los datos.

μ-Argus para la protección de ficheros

El Objetivo

La producción de datos seguros es uno de los principales objetivos perseguidos por las oficinas y agencias de estadística en la fase de difusión de datos. Determinar cuándo un conjunto de datos se puede considerar seguro, no es una tarea trivial y necesita de un criterio de seguridad, previamente determinado, que esté en equilibrio con las reglas y normas que rigen la confidencialidad de los datos y la demanda de información del usuario.

Si el conjunto de datos no es seguro de acuerdo con el criterio impuesto, la información deberá ser modificada de forma que se minimice el riesgo de difusión de datos confidenciales y que, a su vez, mantenga la máxima cantidad de información posible.

μ-Argus es un software orientado hacia la producción de ficheros de datos seguros que basa su funcionamiento en un modelo y un criterio de seguridad propios y que explicaremos brevemente a continuación [22]. También desarrollaremos un ejemplo de aplicación de este software sobre un conjunto de datos pertenecientes a la encuesta de la Población con relación a la Actividad (P.R.A.), elaborada por el EUSTAT y compararemos los ficheros obtenidos como resultado de aplicar distintas codificaciones y transformaciones sobre los datos.

El Modelo

μ-Argus basa su funcionamiento en un modelo de *re-identificación* dónde un hipotético intruso puede identificar a un individuo de la población, vinculando a éste con un determinado conjunto de características recogidas en el fichero de datos. En este modelo juegan un papel de gran importancia las variables más identificativas, a las que denominaremos *variables clave*. Las posibles combinaciones de variables clave que pueden llegar a identificar de forma unívoca a individuos de la población, se denominarán *claves identificativas*.

Así pues, se trata de determinar qué claves identificativas suponen un alto riesgo de difusión de información individual y evitar mediante la *recodificación* o unificación de categorías o mediante *supresión local* de determinados valores de variables, la identificación de individuos o unidades "raros" de la población.

Ser "raro" con respecto a una clave dada, supone que dicha combinación de valores identificativos se da un número de veces menor que un determinado valor prefijado D_k, dónde k es una clave. Si la combinación ocurre más de D_k veces en la población entonces es considerada segura, de lo contrario será insegura y deberán aplicarse las técnicas necesarias para proteger dicha combinación.

El valor D_k será elegido por el encargado de proteger los datos, teniendo en cuenta sus conocimientos previos sobre la población en estudio y lo que se puede considerar como "raro" o no en la misma. En el caso de tener los datos correspondientes a una muestra y no a toda la población, se considerará una estimación de esta frecuencia, de forma que sea proporcional a la fracción de población muestreada.

Técnicas de Protección

Una vez identificadas las claves inseguras, es necesario aplicar las correspondientes modificaciones sobre los datos para que éstos puedan ser publicados con seguridad.

μ-Argus ofrece varios métodos de control de la difusión:

- *La recodificación global de variables*. Se aplica a todos los registros y consiste en la unificación o agregación de categorías.

- *La supresión local* de valores de variables en determinados registros, cuya publicación puede suponer un riesgo de difusión de identidad.

- *Métodos de perturbación* aplicados a variables numéricas y que modifican o transforman los valores de forma que no se pueda conocer su valor exacto (microagregación, aportación de ruido,...).

Muchas veces la aplicación de uno solo de estos métodos no es suficiente para producir un fichero seguro. La combinación de la recodificación de variables junto con la supresión local (intentando que el nº de supresiones sea mínimo), puede aportar un resultado óptimo en términos de pérdida de información y de riesgo de difusión.

Otros Factores de Interés

Medidas de la pérdida de información

μ-Argus utiliza diferentes cuantificadores a la hora de evaluar la pérdida de información según la técnica de protección aplicada. Si lo que se realiza son supresiones locales, el nº de éstas determinará la pérdida de información. Si lo que realizamos es una recodificación global, μ-Argus utiliza una evaluación de la importancia de las variables identificativas (previamente indicada por el protector de los datos en el fichero que describe las variables) así como una evaluación de la importancia de cada código predefinido para cada variable identificativa (el programa permite indicar estas ponderaciones). Estas opciones deben ser indicadas en el caso en el que μ-Argus realiza automáticamente el proceso de producción de datos seguros. Si éste es llevado a cabo interactivamente por el responsable de la producción de datos seguros, puede utilizar cualquier otra medida que considere más adecuada según los datos y según la finalidad de éstos.

Pesos del muestreo

μ-Argus da la opción de aportar ruido al final del proceso de protección, a las variables que contienen las ponderaciones aplicadas en el muestreo. Como ya se vio en el Capítulo 1 del presente cuaderno, estas variables pueden aportar información sobre los estratos de la población y las características que éstos poseen, con el consiguiente riesgo de difusión que esto supone.

Variables Vivienda

Es muy común el muestreo de viviendas a la hora de obtener muestras representativas de la población, entrevistando a los individuos dentro de cada vivienda seleccionada. De esta forma se obtienen variables que presentarán el mismo valor para todos los miembros de una misma vivienda (P.ej. "nº de componentes de la vivienda", "ocupación del cabeza de familia",....). También deberá existir una variable que identifique de forma

unívoca a cada vivienda y que obviamente no podrá ser publicada para su uso externo. Sin embargo esta variable identificativa es utilizada por μ-Argus para reconocer a individuos de una misma vivienda de forma que, si un determinado valor en una variable vivienda debe ser suprimido, lo sea también en el resto de registros que pertenezcan a miembros de la misma vivienda.

Indicadores Regionales

En muchas ocasiones, la publicación de atributos regionales (P.ej. "grado de urbanización", "nº de habitantes", "lugar de trabajo",...) en registros individuales puede aportar pistas sobre las áreas geográficas a las que pertenecen los individuos. Si éstas son áreas muy pequeñas, el riesgo de identificación crece, luego no es deseable que estos indicadores regionales sean publicados en ese caso. μ-Argus, permite trabajar con indicadores regionales y comprueba si la conjunción de éstos supone un riesgo de identificación de pequeñas áreas.

Ejemplo

Para el ejemplo de aplicación del módulo μ-Argus vamos a utilizar la encuesta de Población con relación a la Actividad (P.R.A.) elaborada por el EUSTAT. Tomaremos la muestra para la provincia de Álava recogida en el 1er Trimestre de 1996, lo que supone un total de 2913 individuos.

Para el análisis vamos a utilizar las siguientes variables:

Eciv: Variable categórica que representa el estado civil del individuo. El nivel de identificación asignado a esta variable es 3 y su especificación la siguiente:

0, PNM (Población no muestral)
1, Soltero/a
2, Casado/a
3, Viudo/a
4, Divorciado/a, Separado/a

Edad: Variable que representa la edad del individuo. El nivel de identificación asignado a esta variable es 1. Es decir, es una variable altamente identificativa debido en parte a que se encuentra totalmente desagregada.

Nivel: Variable categórica que representa el nivel de estudios del individuo. El nivel de identificación asignado a esta variable es 3 y su especificación la siguiente:

0, PNM (Población no muestral)
1, Estudios Primarios
2, Estudios Secundarios o FP
3, Estudios Universitarios

Sexo: Variable categórica que representa el sexo del individuo. El nivel de identificación asignado a esta variable es 2. Es decir, se considera muy identificativa aunque el nº de categorías no es grande. Su especificación es la siguiente:

1, Varón
2, Mujer

Prof2: Variable categórica que representa la profesión del individuo. El nivel de identificación asignado a esta variable es 2 y su especificación la siguiente:

0, PNM
1, Técnicos superiores
2, Técnicos medios
3, Personal directivo
4, Jefes administrativos
5, Administrativos
6, Comerciantes/vendedores
7, Auxiliares administrativos
8, Otro personal de servicios
9, Agricultores
10, Siderometalúrgicos
11, Forjadores/fabr.herramientas
12, Mecánicos/montadores
13, Electricistas
14, Fontaneros/soldadores/chapistas
15, Albañiles/obreros de la construcción
16, Conductores y otros obreros del transporte
17, Otros obreros de la industria

Busq1: Variable categórica que indica la situación del individuo con respecto a la búsqueda de empleo. El nivel identificativo asignado a esta variable es 0 y su especificación la siguiente:

0, PNM (Población no muestral)
1, Busca empleo
2, No busca

Pra1: Variable categórica que representa la situación del individuo con respecto al empleo. El nivel identificativo asignado a esta variable es 0 y su especificación la siguiente:

0, PNM (Población no muestral)
1, Ocupados
2, Parados que ya han trabajado
3, Parados que buscan su 1er.empleo
4, Inactivos

Tjor: Variable categórica que representa el tipo de jornada laboral que lleva a cabo el individuo. El nivel identificativo asignado a esta variable es 0 y su especificación la siguiente:

0, PNM (población no muestral)+inactivos+parados
1, Completa
2, Parcial

Htrt: Variable numérica que representa las horas totales de trabajo a la semana que realiza el individuo.

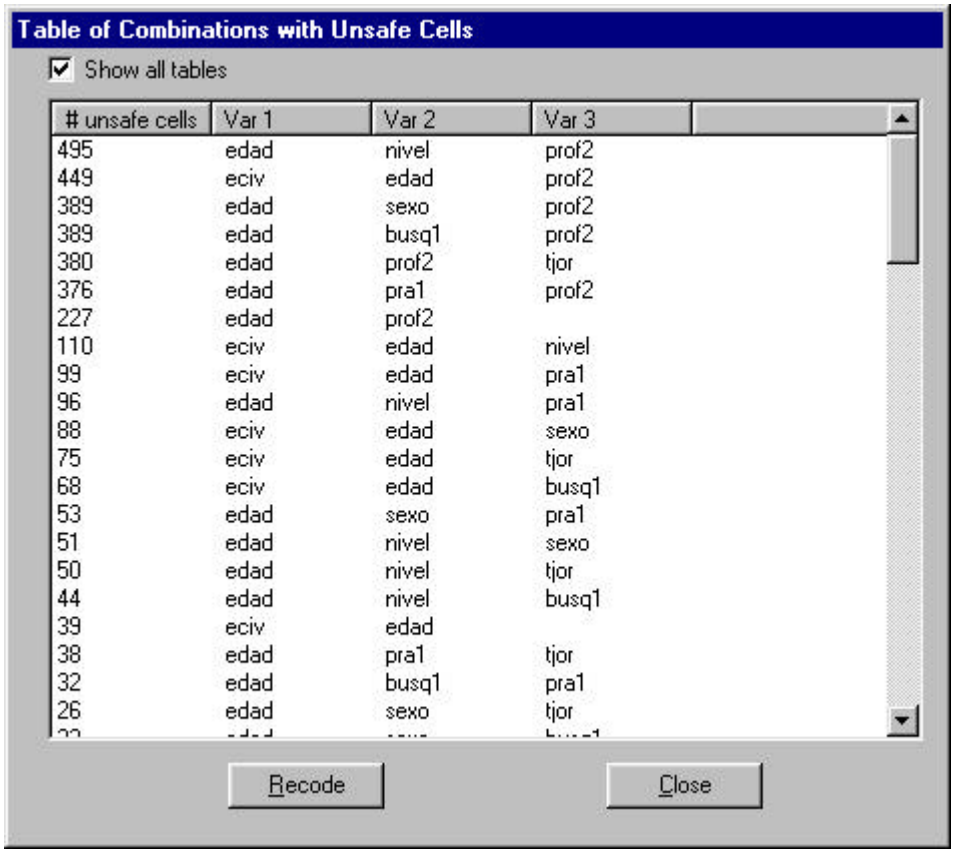
Dpar1: Variable numérica que representa la duración del paro en meses para el individuo.

Eleva: Variable peso que representa las ponderaciones que corresponden a cada individuo dentro de la población.

Vamos a realizar el estudio de las claves identificativas de dimensión 3. Es decir, el programa creará todas las posibles tablas de frecuencias que se originan del cruce de 3 variables (en nuestro caso con todas las variables categóricas) y contará el nº de celdas inseguras para cada combinación de variables.

Para este ejemplo consideraremos que una celda será insegura si es unitaria (D_k=1), es decir, vamos a determinar el nº de individuos que son únicos con respecto a una determinada clave. De esta forma y observando las claves que generan un mayor nº de celdas unitarias, podremos decidir sobre que variables debemos realizar las recodificaciones y transformaciones necesarias para reducir el nº de celdas inseguras.

Esta es la pantalla de información que presenta μ-Argus una vez generadas todas las posibles combinaciones de 3 variables:



# unsafe cells	Var 1	Var 2	Var 3
495	edad	nivel	prof2
449	eciv	edad	prof2
389	edad	sexo	prof2
389	edad	busq1	prof2
380	edad	prof2	tjor
376	edad	pra1	prof2
227	edad	prof2	
110	eciv	edad	nivel
99	eciv	edad	pra1
96	edad	nivel	pra1
88	eciv	edad	sexo
75	eciv	edad	tjor
68	eciv	edad	busq1
53	edad	sexo	pra1
51	edad	nivel	sexo
50	edad	nivel	tjor
44	edad	nivel	busq1
39	eciv	edad	
38	edad	pra1	tjor
32	edad	busq1	pra1
26	edad	sexo	tjor

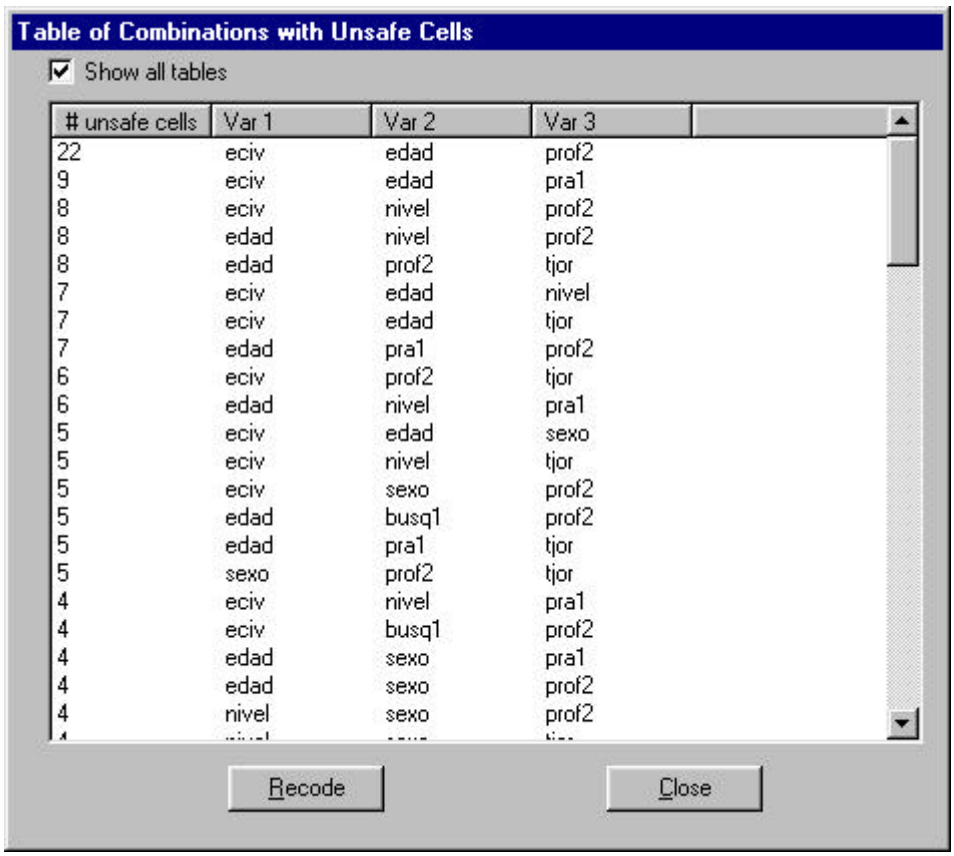
Se observa que la clave que mayor nº de celdas unitarias, y por lo tanto inseguras, genera es *edad x nivel x prof2* y que las variables que más aparecen en las claves con mayor nº de celdas inseguras son *edad* y *prof2*. Con la ayuda del programa

realizaremos una *recodificación global* de estas dos variables para reducir el nº de celdas inseguras. Las nuevas codificaciones para las variables serán las siguientes:

Edad: 1, < 16 años
2, [16-24]
3, [25-34]
4, [35-44]
5, [45-54]
6, [55-64]
7, > 64 años

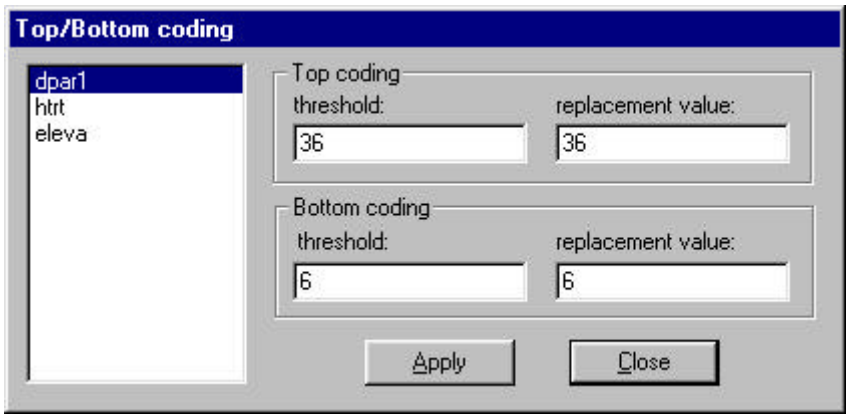
Prof2: 0, PNM
1, Profesionales y Técnicos
2, Personal Directivo
3, Empleados administrativos
4, Comerciantes y vendedores
5, Personal de servicios
6, Agricultores
7, Obreros

Una vez realizada esta recodificación observamos que el nº de celdas inseguras se ha reducido notablemente.



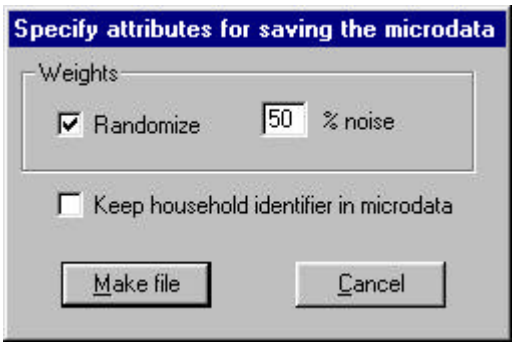
El resto de celdas inseguras se protegerán mediante *supresiones locales* que el programa realiza automáticamente cuando crea el fichero seguro.

Antes de terminar el proceso podemos aplicar alguno de los procedimientos de protección para variables numéricas que aporta µ-Argus. Para nuestro ejemplo aplicaremos un Top-Bottom coding a la variable *dpar1:*

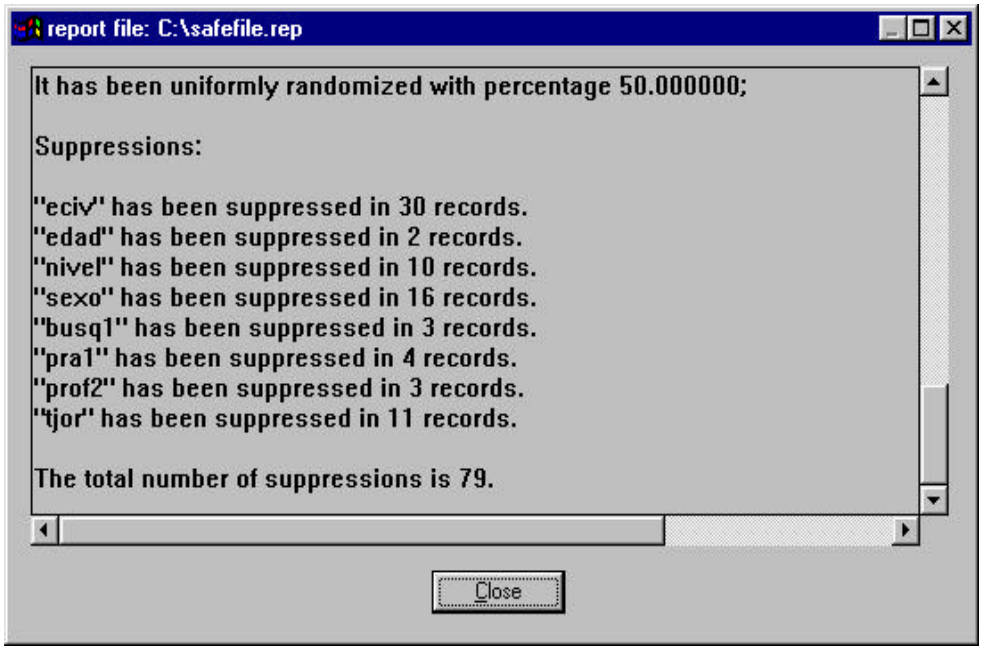


Es decir, para todos los parados de larga duración (más de 36 meses en paro) el valor de la variable *dpar* será sustituído por el valor de reemplazamiento (en este caso 36) y para todos aquellos que lleven menos de 6 meses en el paro el valor de la variable *dpar* será 6 (valor de reemplazamiento).

Por último sólo nos queda crear el nuevo fichero de datos con todas las recodificaciones y transformaciones realizadas sobre las variables. El programa nos permite en este momento la posibilidad de aleatorizar la variable *Eleva* (pesos muestrales) en un determinado porcentage:



Cuando creamos el fichero seguro, Argus realiza las supresiones locales necesarias para evitar la identificación de las celdas unitarias minimizando siempre la pérdida de información. Es decir, si para un mismo registro tenemos más de una clave insegura y estas claves tienen alguna variable en común, el valor suprimido será el de la variable común. Si no es así, Argus suprimirá el valor de aquella variable con un nivel de identificación menor. Para nuestro ejemplo estas son las supresiones que Argus realiza automáticamente:



De un fichero original con un total de 32043 datos (11 variables x 2913 registros) hemos obtenido un fichero seguro con respecto a claves de 3 dimensiones, donde tan solo hemos suprimido 79 datos.

Es posible proteger conjuntos de datos con respecto a claves de dimensiones mayores o con respecto a combinaciones de variables concretas que pueden suponer un alto riesgo de identificación para los datos que se manejan o para la población en estudio.

La elección de una u otra clave o conjunto de claves a analizar, debe ser realizadas por la persona o personas encargadas de la protección de los datos, de los que se asume un determinado conocimiento "a priori" de la población en estudio y de lo que en ella supone un mayor o menor riesgo de difusión.

Capítulo

4

Protección en Tablas

Granularidad

El concepto de granularidad es explicado en [14] utilizando una analogía entre la observación de una tabla de datos y una imagen fotográfica. Si ampliamos indefinidamente una fotografía, lo único que observaremos serán puntos o "granos", perdiendo definición y con ello la visión global de la imagen. El observador se centra en el proceso de producción de la imagen, en el detalle. Lo mismo ocurre en una tabla de datos dónde el fin último es informar sobre características más o menos agregadas de la población (imagen global) y no sobre características detalladas de los individuos (imagen ampliada). Si esta tabla contiene gran cantidad de celdas unitarias (sólo un individuo contribuye a la celda), el foco de atención se desvía hacia respuestas individuales con el consiguiente riesgo de identificación. El nivel de "granularidad" de una tabla nos ayudará a reconocer si ésta es o no segura y si es necesaria la aplicación de medidas de protección para evitar la difusión de información individual.

Definición y Cuantificación

Los datos agregados en forma tabular no nos proporcionan datos exactos de cada individuo o unidad de la población, pero sí informa sobre las contribuciones de dichos individuos a las celdas y su distribución para una determinada fila o columna de la tabla.

Al contrario que en el caso de conjuntos de datos almacenados en ficheros de registros donde la difusión de un único registro amenaza la seguridad de los datos, el riesgo de difusión en tablas vendrá dado por la acumulación de un determinado número de celdas unitarias en una determinada *hiperfila.*(*)

Podemos entender el concepto de "granularidad" como un *criterio de seguridad para tablas de datos agregados, basado en el número de celdas unitarias o "granos" que contiene y su distribución con respecto a las demás celdas de su misma hiperfila.*

Consideraremos que una tabla es o no segura si sobrepasa un determinado *ratio de granularidad* definido para cada hiperfila como:

$$Ratio\ de\ Granularidad = \frac{n^o\ de\ celdas\ unitarias}{n^o\ total\ de\ celdas\ con\ valor} \times 100$$

Si este ratio está por debajo del 50% no se toma ninguna medida sobre la hiperfila, en otro caso sí. Un posible procedimiento de decisión sobre una tabla deberá tener en cuenta la incidencia de la granularidad sobre todos las posibles hiperfilas generadas fijando categorías en variables identificativas.

(*) Subconjunto de celdas, originadas por el cruce de variables identificativas y definidas fijando categorías en una o varias de ellas. P.ej. Ocupación x Región x Sexo definen una hiperfila para la categoría Sexo=Mujer y Ocupación=Estadístico.

Atributos Cualitativos de Granularidad

Aparte del aspecto cuantitativo de la granularidad, ciertas características cualitativas influyen también en ésta, afectando al riesgo de difusión de datos:

- *Grado de estratificación de las observaciones.* Cuanto mayor sea el grado de estratificación, más desagregada se encontrará la población distribuida a lo largo de una determinada hiperfila y mayor será el riesgo de unicidad de las celdas.

- *Patrones Inesperados.* Una hiperfila puede ser vista como un vector de frecuencias que define la distribución muestral de una subpoblación. Si estas frecuencias son "insólitas" con respecto a la distribución de la población (es decir, no se adaptan al patrón poblacional), esto nos puede estar indicando un alto índice de granularidad.

Tipos de Granularidad

Para definir los tipos de granularidad vamos a utilizar un ejemplo de hiperfila dentro de una hipotética tabla tridimensional Ocupación x Edad x Sexo, tabulada para una población de tamaño medio y situada cerca de un gran núcleo administrativo.

- La siguiente hiperfila representaría una situación normal, dónde el índice de granularidad es bajo y por lo tanto también lo es el riesgo de difusión:

Normal.- No existe Granularidad

Sexo=Mujer	EDAD				
	5-19	20-34	35-50	50-65	66+
...					
Ocupación=Estadístico	1	30	18	9	2
...					

La celda unitaria en el grupo de edad 5-19 no añade riesgo alguno de identificación ya que la distribución de la hiperfila se adapta a la esperada por las características de la población.

- *Granularidad Proporcional*. Es aquella que es mayor o igual que un determinado nivel crítico o ratio de granularidad (en nuestro caso definido por 0.5).

Proporcional.- Ratio ³ 0.5

Sexo=Mujer	EDAD				
	5-19	20-34	35-50	50-65	66+
...					
Ocupación=Estadístico	1	3	4	1	1
...					

- *Granularidad "Rara" o patrón inusual*. Existen celdas que son insólitamente unitarias y que no se esperaban que lo fueran por las características de la población o por la distribución a priori de la misma.

"Rara".- Patrón Inesperado

Sexo=Mujer	EDAD				
	5-19	20-34	35-50	50-65	66+
...					
Ocupación=Estadístico	0	1	1	21	1

En este caso sería de esperar una frecuencia mucho más alta en el grupo de edad 20-34 y lo mismo para el 35-50.

- *Granularidad Dispersa*. Contiene celdas unitarias aisladas entre celdas con frecuencias muy altas.

Dispersa.

Sexo=Mujer	EDAD				
	5-19	20-34	35-50	50-65	66+
...					
Ocupación=Estadístico	1	30	1	9	1
...					

En base a los distintos tipos de granularidad definidos, se pueden crear los correspondientes patrones críticos de granularidad. Si una tabla presenta alguno de estos patrones, esto nos alertará sobre la existencia de un alto índice de granularidad por lo que se pueden llevar a cabo acciones contra un posible riesgo de difusión de información individual.

Aplicación en Microtablas Electrónicas

El concepto de Microtabla hace referencia a la *difusión pública mediante medios electrónicos, de tablas multidimensionales y altamente desagregadas.* Son usadas para modelizar poblaciones ricas en estructura y resultan realmente útiles a la hora de realizar estudios exploratorios de la población, en casos en los que la publicación mediante ficheros de registros supondría un alto riesgo de difusión de datos confidenciales.

El riesgo de identificación en Microtablas vendrá dado por la distribución de contribuciones unitarias a las celdas. Aquí es donde adquiere sentido el concepto de granularidad y su aplicación como detector de tablas "no seguras".

Una Microtabla se comporta como un fichero de registros dónde los campos son cada una de las variables de la tabla y las "unidades" o registros están definidos por los valores de las celdas para cada hiperfila. La equiparación de una tabla a un conjunto de registros, convierte los procedimientos de protección de ficheros de datos *(Ver Capítulo 2)*, en potenciales técnicas de detección de granularidad.

Este es el caso de μ-Argus, cuyo funcionamiento basado en la detección de individuos "raros" (únicos en la población), puede adaptarse a la búsqueda de celdas unitarias dentro de microtablas. El siguiente algoritmo es un prototipo de implementación que combina la detección de granularidad en tablas con un procedimiento de control específico para ficheros de registros como es μ-Argus.

Algoritmo para la eliminación de granularidad en ficheros de datos

Entrada. Fichero de datos con registros confidenciales.

Salida. Fichero de datos protegidos mediante criterios de granularidad.

Paso 1. μ-Argus genera tablas bidimensionales y tridimensionales para las n variables identificativas especificadas.

Paso 2. Si la proporción de celdas unitarias a lo largo de cualquier hiperfila dada (construida fijando los valores de las variables identificativas), es ≥ 0.5, entonces la tabla completa es considerada insegura.(*)

Paso 3. Si la tabla es insegura, se redefinen las filas o columnas afectadas mediante agregación de categorías o supresión de algunas celdas unitarias.

Paso 4. Se repite el paso 3 hasta que la granularidad descienda por debajo de un determinado nivel de seguridad, para cada hiperfila y tabla inseguras.

Paso 5. Se reconstruye el fichero de datos protegido de acuerdo a criterios de granularidad y producido vía μ-Argus.

La aplicación de criterios de granularidad para la creación de ficheros de datos seguros, evita en ocasiones la excesiva pérdida de información a la que llevan muchas veces los métodos de supresión de celdas, para los que no se han encontrado todavía procedimientos que proporcionen la solución más óptima, acercándose en todo caso a ésta última.

Por otro lado un si un fichero generado de esta forma es seguro, también lo será cualquier tabla que se derive de él, por lo que se provee de un método para la creación de ficheros que pueden ser destinados a uso externo (en bases de datos externas), dónde el usuario construye las tablas que le son de utilidad, a la medida de sus necesidades.

Método de Redondeo

Se trata de un método sencillo y fácil de implementar, utilizado tanto en tablas de frecuencias como de magnitud, de dimensiones no muy altas (de doble o triple entrada). Esta técnica permite elegir una *base* entera, sobre la cual se realiza el redondeo, siempre de una forma controlada, ya que el redondeo de tablas puede afectar a la aditividad de los totales o subtotales.

Definición 1. *Dada una base de redondeo entera b, el procedimiento de redondeo de la tabla A consiste en sustituir cada frecuencia o total de la tabla por uno de los dos múltiplos enteros de b más próximos al valor de la celda. Si este valor ya es un múltiplo de b, se tomarán como valores adyacentes, el de la propia celda y el siguiente múltiplo más grande de b.*

(*) Se puede aplicar cualquier otro índice de granularidad tanto cualitativo como cuantitativo si conocemos el tipo de granularidad que afecta a la población.

En el caso de tablas bidimensionales, el redondeo se puede restringir a condiciones de no-nulidad, esto supone que si la entrada es ya un múltiplo de b, se mantiene su valor fijo.

El problema del paso a tres o más dimensiones, no es fácil de resolver. En teoría, las tablas pueden ser de cualquier dimensión aunque en la práctica las tablas de dos y tres dimensiones son las más comunes. En Ernst(1989)(*) se propone una solución para tablas de tres dimensiones basado en un *redondeo sucesivo*, es decir redondear sobre el valor redondeado. Se trata de un método eficiente aunque no muy extendido y además respeta las condiciones de no nulidad anteriormente citadas.

Sistemas de Supresión de Celdas

Los sistemas de supresión de celdas son uno de los mecanismos de protección de tablas más extendidos y comúnmente usados por las oficinas y agencias de estadística. Se aplican generalmente en *tablas de magnitudes agregadas* aunque pueden adaptarse también al caso de *tablas de frecuencias.* Estas técnicas no resultan tan sencillas de automatizar como el redondeo ya que no sólo es necesaria una supresión de las celdas confidenciales o sensitivas, sino que precisa de una *supresión secundaria* de las celdas que pueden aportar información sobre las eliminadas con el primer patrón de supresiones.

Actualmente la supresión primaria de celdas se realiza automáticamente, sin embargo la segunda parte del proceso se realiza en muchos casos de forma manual. En este apartado trataremos de explicar las claves para la automatización de un sistema eficiente para la supresión secundaria de celdas y abordaremos la explicación de un modelo de programación lineal entera que nos aportará el patrón óptimo de supresiones en términos de pérdida de información.

Medidas de Sensitividad

Antes de afrontar una supresión de celdas, hay que determinar cuáles de ellas suponen un riesgo de difusión en base a unas determinadas *reglas de sensitividad*. La aplicación de una u otra regla afectará posteriormente al patrón de supresiones tanto primario como secundario. Algunas de estas reglas son:

- **Regla (n, k).** Una celda será suprimida si las n contribuciones más grandes al valor de la celda constituyen más del k% del valor total de la misma.

- **Regla del porcentaje p.** Una celda deberá ser suprimida si un usuario es capaz de aproximar la contribución de un determinado individuo al valor de la celda en aproximadamente un $\pm$ p%.

- **Regla p-q.** Es una extensión de la anterior, en dónde son tenidos en cuenta los conocimientos previos que el usuario tiene de la población, medidos de nuevo en porcentajes de la contribución de un individuo a una celda.

Todas las reglas anteriores se basan en la contribución de los individuos al valor de las celdas. Si una celda es "dominada" por uno o dos individuos, más protección requerirá.

(*) Ernst, L.R.(1989) "Further applications of linear programming to sampling problems". Technical Report-Census SRD (RR-89-05)

Medidas de la Pérdida de Información

Obviamente, lo que se pretende al aplicar cualquier patrón de supresiones, es minimizar la pérdida de información. Para ello se especifica una función de esta pérdida que puede venir dada por:

- **El número de celdas suprimidas.** La optimización de esta función llevará a la supresión de un menor número de celdas de valor grande.

- **La suma de los valores de las celdas suprimidas.** Para minimizar esta función se preferirá eliminar un mayor número de celdas con valores pequeños.

Según el uso que se le dé a la tabla, deberemos escoger una función de pérdida u otra, obteniendo para cada caso un patrón de supresiones distinto. Muchas veces este patrón ha de ajustarse a las necesidades de un usuario o cliente que demanda la información, por lo que el sistema ha de permitir "marcar" determinadas celdas que pueden ser de mayor utilidad o importancia para el usuario, de forma que la probabilidad de que éstas sean incluidas en el patrón de supresiones sea mínima.

Un modelo de Programación Lineal Entera Mixta para la Supresión Secundaria de Celdas **[8]**

Para introducir el problema veremos un ejemplo sencillo para una tabla de dos dimensiones, utilizado por Willenborg y Waal:(*)

Ejemplo

Las siguientes tablas muestran las inversiones realizadas por empresas (en millones de unidades monetarias) por región y actividad en un determinado periodo:

	A	B	C	Total
Actividad I	2	50	10	80
Actividad II	8	19	22	49
Actividad III	1	32	12	61
Total	4	10	44	190

(a) Tabla original

	A	B	C	Total
Actividad I	20	50	10	80
Actividad II	*	19	*	49
Actividad III	*	32	*	61
Total	45	10	44	190

(b) Tabla publicada

Asumimos que la información correspondiente a la Actividad II en la región C es confidencial, luego la celda con valor 22 se considera sensitiva y no debe ser publicada. Sin embargo esto no resulta suficiente ya que, debido a la existencia de totales y subtotales, es posible recalcular su valor. Así pues, es necesario suprimir otras entradas de la tabla (véase b). De esta forma el valor exacto de la celda confidencial no puede ser calculado, aun cuando es posible determinar un rango de valores posibles para la celda, consistentes con los publicados.

Si denominamos $y_{23}(inf)$ al valor mínimo que puede tomar la celda sensitiva, éste puede ser calculado resolviendo el siguiente problema de programación lineal, en el cual las incógnitas y_{ij}' representan a los valores suprimidos (i,j) de la tabla:

(*) Willenborg and Waal, "Statistical Disclosure Control in Practice", Lecture Notes in Statistics 111, Springer, New York, 1996.

$$
\begin{aligned}
& y_{23}\,(\inf) \;=\; min \quad y_{23}' \\
& s.a. \\
& \quad y_{21}' + y_{23}' = 30 \\
& \quad y_{31}' + y_{33}' = 29 \\
& \quad y_{21}' + y_{31}' = 25 \\
& \quad y_{23}' + y_{33}' = 34 \\
& y_{21}', y_{23}', y_{31}', y_{33}' \geq 0
\end{aligned}
$$

De la misma manera se puede calcular el valor máximo $y_{23}(sup)$ cambiando la función objetivo y sujeto a las mismas restricciones.

En este caso las soluciones son $y_{23}(inf)=5$, $y_{23}(sup)=30$ y podemos decir que la información sensitiva está protegida dentro del intervalo [5,30]. Si este intervalo es considerado lo suficientemente amplio por la agencia o instituto de estadística, entonces la celda se considerará protegida. A veces se pueden aplicar las denominadas *cotas externas* para estrechar este intervalo. Éstas son calculadas suponiendo que la variación del valor nominal de la celda sensitiva no será superior a un cierto porcentaje. P.ej. Si en nuestro caso acotamos la variación del valor sensitivo en un ±50%, es decir añadimos la condición $11 \leq y_{23}' \leq 33$ al problema de programación lineal, obtenemos el intervalo de protección [18,26] más acorde con la realidad.

Notación:

Dado un conjunto de celdas SP *(supresiones primarias)* junto con los niveles de protección requeridos para cada una (fijados por al agencia o el instituto de estadística), el objetivo final consiste en calcular el conjunto de supresiones secundarias óptimo que proteja a todas las celdas sensitivas de un posible atacante, y de forma que la pérdida de información que suponen dichas supresiones sea mínima.

Sean:

$i = \{1,.....,n\} \rightarrow$ conjunto de índices de la tabla a proteger.

$w_i \geq 0$ el coste de supresión de la celda i

$[a_i] \rightarrow$ valores para las entradas de la tabla original *(tabla nominal)*

$[y_i] \rightarrow$ valores factibles para las entradas de la tabla *(tabla factible)*

Decimos que un vector $[y_i]$ define una tabla factible cuando Ay=b, dónde A es una matriz {0,1,-1} y b=0

Asumimos que el atacante conoce un rango de valores posibles (cotas externas) para cada entrada de la tabla a_i , es decir $[a_i\text{-}lb_i, a_i\text{+}ub_i]$ y se cumple $a_i - lb_i \leq y_i \leq a_i + ub_i$ para todo i. .

Sean $i_1,.....i_p$ con $p = |SP|$, los índices del conjunto de celdas sensitivas SP.

Si $[a_{ik} - LPL_k,\ a_{ik} + UPL_k] \subset [y_i(inf),\ y_i(sup)]\ \forall\ k=\{1,.....,p\} \Rightarrow$ tabla segura

Dónde LPL_k y UPL_k son los niveles inferior y superior de protección requeridos por la agencia o instituto de estadística y $[y_i(inf), y_i(sup)]$ es el intervalo calculado por un posible atacante.

El Modelo

Dado un conjunto SUP de supresiones (primarias y secundarias), el problema de un supuesto atacante consiste en calcular valores máximos y mínimos para las celdas incluidas en SP dentro de una tabla dónde sólo los valores publicados coinciden con los originales de la tabla nominal. Para cada i_k tendremos asociado un subproblema del tipo:

$$
\left.
\begin{array}{l}
y_{i\,k}(inf) = \min y_{i\,k}' \\
\qquad Ay' = b \qquad\qquad (1) \\
a_i - lb_i \le y_i' \le a_i + ub_i \quad \text{para } i = 1,\ldots,n \\
\qquad y_i' = a_i \qquad \text{para todo } i \notin SUP
\end{array}
\right\}
$$

$$
\left.
\begin{array}{l}
y_{i\,k}(sup) = \max y_{i\,k}'' \\
\qquad Ay'' = b \qquad\qquad (2) \\
a_i - lb_i \le y_i'' \le a_i + ub_i \quad \text{para } i = 1,\ldots,n \\
\qquad y_i'' = a_i \qquad \text{para todo } i \notin SUP
\end{array}
\right\}
$$

Dónde las variables continuas y' e y'' son locales para cada subproblema.

Para poder formular lo anterior como un modelo de programación lineal entera mixta introducimos una variable binaria X definida de la forma siguiente:

$$
\left.
\begin{array}{l}
x_i = 1 \quad \text{si } i \in SUP \quad \text{(celda suprimida)} \\
x_i = 0 \quad \text{en otro caso}
\end{array}
\right\}
$$

Así obtenemos la siguiente función a minimizar:

$$\min \sum_{i=1}^{n} w_i x_i$$

Sujeto a:

$x \in \{0,1\}^n$ y para cada $i_k \in SP$, $x_{i\,k} = 1$

$$LPL_k - a_{ik} \leq -y_{ik}(inf) = \max(-y_{ik}')$$
$$Ay' = b \qquad (3)$$
$$a_i - lb_i x_i \leq y_i' \leq a_i + ub_i x_i \quad \text{para } i = 1,.....,n$$

$$LPL_k - a_{ik} \leq -y_{ik}(sup) = \max y_{ik}''$$
$$Ay'' = b \qquad (4)$$
$$a_i - lb_i x_i \leq y_i'' \leq a_i + ub_i x_i \quad \text{para } i = 1,.....,n$$

Para resolver un problema tan complejo, dónde existe un gran número de variables locales (y' e y'') e interrelaciones entre éstas y la variable introducida X, es necesaria la "relajación" del modelo (es decir, permitir que la condición $x_i \in \{0,1\}$ se convierta en $(0 \leq x_i \leq 1)$. Además plantearemos un modelo de programación lineal, que denominaremos *problema master*, dependiente inicialmente sólo de las variables x_i:

$$min \sum_{i=1}^{n} w_i x_i : x_{i_1} = x_{i_2} = ... = x_{i_p} = 1,\ x \in [0,1]^n$$

Sea x^* la solución óptima encontrada para el anterior problema. Ésta es sustituida en (3) y (4) en lo que denominamos el *procedimiento de verificación,* que consiste en la resolución de los 2p pequeños problemas de programación lineal, que devolverán como resultado los vectores locales y' e y'' que cumplen las condiciones impuestas por los límites de protección y confirmen la optimalidad de x^*. Si x^* es no factible, el procedimiento habrá encontrado, para algunos de los problemas, algún valor estrictamente menor que el de los límites de protección impuestos $LPL_k - a_{ik}$ o $LPL_k + a_{ik}$, y para cada solución dual de dichos problemas se puede derivar una desigualdad lineal (dependiente sólo de x) que no es cumplida por x^*, luego certifica que x^* no es una posible solución del problema completo.

Estas nuevas desigualdades, denominadas *cotas de Benders*, son añadidas al problema master y éste es de nuevo optimizado (mediante técnicas paramétricas, p.ej algoritmo de simplex). El proceso se itera hasta encontrar una solución óptima x^* para el modelo relajado. Si esta solución es entera, entonces será solución del problema inicial de programación lineal entera. Si no es el caso, hay que aplicar métodos de ramificación-acotación para forzar la integridad de x^*.(*)

Las cotas de Benders calculadas durante este proceso son de la forma:

$$\sum_{i=1}^{n} (b_i' ub_i - a_i' lb_i) x_i \geq LPL_k$$
$$\sum_{i=1}^{n} (b_i'' ub_i - a_i'' lb_i) x_i \geq LPL_k$$

(*) M.Padberg, G,Rinaldi, "A branch-and-cut algorithm for the resolution of large scale symetric traveling salesman problems", SIAM Reviews,33 (1991)

Dónde (α',β') y (α",β") son soluciones duales óptimas para los problemas (3) y (4).

Estas cotas, llamadas también *cotas de capacidad*, pueden ser reforzadas reemplazando $(b'_i\ ub_i - a'_i\ ub_i)$ por $min\{(b'_i\ ub_i - a'_i\ ub_i), LPL_k\}$ y $(b"_i\ ub_i - a"_i\ ub_i)$ por $min\{(b"_i ub_i - a"_i\ ub_i), LPL_k\}$, lo que resulta muy efectivo en la práctica a la hora de reducir el número de iteraciones del proceso.

Otras Claves de importancia

Otros factores a tener en cuenta al ahora de implementar un sistema eficiente de supresión de celdas son los siguientes:

- **Tratamiento de los "ceros".** Una celda con valor 0, nos está proporcionando una información exacta sobre las características que unos determinados individuos **no** tienen, lo que también supone un riesgo de difusión si estos datos son confidenciales. Cómo proteger una celda con contribuciones nulas no es un problema trivial. Utilizando cualquiera de las reglas de sensitividad anteriormente citadas, nunca se consideraría una celda nula como sensitiva (evidentemente, no existe ninguna contribución dominante al valor de la celda), por lo que necesitan un tratamiento especial que el sistema de supresiones ha de tener en cuenta.

- **Supresión Simultánea.** Un sistema de supresión simultánea busca el mejor patrón de supresiones que proteja "de una sola vez" a todas las celdas sensitivas. Sin embargo, es solamente aplicable en casos de tablas muy sencillas, ya que resulta poco eficiente en tiempo debido a que chequea todos los patrones de supresión posibles para una tabla.

- **Supresión Secuencial.** Protege las supresiones primarias, secuencialmente, asegurando la protección en cada paso del proceso. Este sistema tiende a sobreproteger las celdas si no se le dota de una "memoria" que recoja las celdas que van siendo incluidas en el patrón de supresión. De esta manera el sistema puede utilizar en cada paso celdas suprimidas anteriormente de forma que se eviten supresiones que podrían resultar redundantes o innecesarias.

- **Múltiples Dimensiones.** La protección en tablas de varias dimensiones ha encontrado en los métodos de Programación Lineal una solución eficaz y válida. Aun cuando no siempre aporta la solución óptima, sí se aproxima a ella y en muchos casos puede ser "refinada" mediante la aplicación de métodos heurísticos.

- **Múltiples Tablas.** La única forma de garantizar una protección eficaz entre tablas consiste en producir todas ellas a partir de un único conjunto de datos. El sistema de supresión ha de ser capaz de detectar celdas comunes en múltiples tablas que muchas veces no se producen de una forma simultánea en el tiempo. Esto supone el mantenimiento de una memoria histórica de todas las supresiones realizadas en tablas producidas en el pasado, con lo que esto requiere en capacidad de control y almacenamiento.

τ- Argus para la protección de tablas

τ-Argus es el módulo específico para la protección de tablas integrado en el paquete de software ARGUS [23]. Nos va a permitir crear tablas "seguras" a partir de ficheros de datos planos aplicando las técnicas más utilizadas en la protección de datos tabulares.

Los principales métodos que va a utilizar τ-Argus en el proceso van a ser el *redondeo controlado* y la *supresión de celdas* (tanto primaria como secundaria). Para ambos utilizará el modelo de programación lineal entera que hemos especificado en el apartado anterior. También permitirá la recodificación global de las variables antes de aplicar las técnicas mencionadas, de forma que se reduzcan al máximo el nº de celdas sensitivas o confidenciales generadas por la aplicación del criterio de sensitividad (en este caso τ-Argus aplica la regla de sensitividad (n,k)).

¿Cómo funciona τ-Argus?

Fases del funcionamiento de τ-Argus:

1. Lectura del fichero de datos y del fichero de descripción de variables.
2. Especificación de la tabla a proteger (hasta 3 dimensiones)
3. Indicación de los criterios de sensitividad.
4. Creación de la tabla y determinación del nº de celdas sensitivas o confidenciales.
5. Aplicación del método de recodificación global hasta minimizar el nº de celdas sensitivas
6. Aplicación el procedimiento de supresión secundaria de celdas, redondeo o ambas, para proteger a las celdas sensitivas.
7. Creación y almacenamiento de la tabla segura.

Ejemplo

A continuación desarrollaremos un ejemplo de aplicación con datos sobre una encuesta económica elaborada por el EUSTAT. La población está formada por 313 empresas que desarrollan I+D en Euskadi.

1. Los datos corresponden al año 1996 y las variables que se van a utilizar en el proceso son las siguientes:

act18: Variable categórica que representa la actividad a la que se dedica la empresa clasificada en 18 ramas de actividad.

tamaño: Variable categórica que representa el tamaño de la empresa en nº de empleados dividida en 7 categorías de tamaño.

gastot: Variable numérica que representa los gastos financieros totales soportados por la empresa en millones de pesetas.

2. Vamos a especificar la tabla bidimensional *act18 x tamaño*. Crearemos la tabla de frecuencias (nº de empresas que contribuyen a cada celda) y a continuación una tabla de magnitud para el mismo cruce de variables categóricas, representando la variable *gastot* como ítem dentro de cada celda.

3. Como criterios de sensitividad indicamos los siguientes:

- Para la tabla de frecuencias serán consideradas como celdas sensitivas aquellas de valor ≤ 2. (dos empresas o menos contribuyen a la celda)

- Para la tabla de magnitud serán consideradas como sensitivas aquellas celdas en las que 2 o menos empresas contribuyen al valor de la misma y todas aquellas cuya contribución dominante suponga más de un 75% del valor total de la celda.

4. Veamos a continuación las tablas que crea Argus y que celdas considera sensitivas según los criterios especificados:

tamaño act18	<25	25-49	50-99	100-249	250-499	500-999	>=1000	total
Agropecuario y pesca	**2**	.	.	.	.	.	.	**2**
Química	7	3	6	**2**	5	**1**	.	24
Caucho plástico	**2**	.	3	8	3	.	**1**	17
Metalurgia	.	**2**	**1**	7	6	**2**	**2**	20
Artículos metálicos	5	.	7	6	8	4	**1**	31
Máquina herramienta	.	4	9	4	**2**	**1**	.	20
Aparatos domésticos	.	.	**2**	**2**	**2**	**1**	**1**	8
Otra maquinaria	**1**	4	13	11	3	**1**	.	33
Material eléctrico	5	6	**2**	3	5	.	.	21
Material electrónico	4	**2**	**2**	**2**	**2**	.	.	12
Material de precisión	12	**2**	3	**2**	**1**	.	.	20
Material de transporte	.	**2**	**2**	4	6	.	**2**	16
Otras manufacturas	3	4	7	5	**2**	.	.	21
Energía y Construcción	**1**	**1**	**2**	**3**	**1**	.	**2**	10
Actividades Informáticas	5	**2**	3	**1**	.	.	.	11
Actividades I+D	6	3	3	6	.	.	.	18
Otras actividades empr.	14	5	3	**1**	**2**	**1**	.	26
Otros Servicios	**2**	.	.	**1**	.	.	.	3
Total	69	40	68	68	48	11	9	313

Las celdas señaladas en negrita son las sensitivas o confidenciales de la tabla de frecuencias. Como podemos comprobar el ejemplo elegido es muy sensitivo ya que la población no es muy grande (313 empresas) y está muy desagregada. Tendremos que recodificar ambas variables con el fin de disminuir el nº de celdas sensitivas.

Para la tabla de magnitud (los ítems dentro de la celda corresponden a la variable numérica *gastot*) tendremos el mismo nº de celdas sensitivas + aquellas que cumplan el criterio (n,k) impuesto.

5. Recodificamos las variables *act18* y *tamaño* de la siguiente forma:

act18 pasará a tener 9 categorías

tamaño se recodificará en 2 categorías (empresas con <100 empleados y resto)

La tabla de magnitud que Argus crea con estas nuevas recodificaciones es la siguiente (las celdas sensitivas están señaladas en negrita)

Gastos financieros totales (gastot)

Tamaño act18	<= 100 trabajadores	>100 trabajadores	Total
Agropecuario y Pesca	**22196**	.	**22196**
Química	593301	1425519	2018820
Caucho plástico	388895	486771	875666
Metal	834529	2962455	3796984
Maquinaria	2274677	2740140	5014817
Materiales y manufacturas	4032117	8563599	12595716
Energía y construcción	207072	350013	557085
Otras actividades	16535820	**1726625**	18262444
Otros servicios	**171819**	.	**171819**
Total	25060426	18255122	43315548

El nº de celdas sensitivas se ha reducido de forma considerable. Las celdas en negrita no serán publicadas junto con alguna supresión complementaria que proteja a aquellas de ser recalculadas.

6. Argus nos proporciona un patrón de supresiones secundarias que protege a las confidenciales dentro de unos límites de seguridad impuestos por la agencia o instituto de estadística. En nuestro caso 70% y 130% (es decir, protegemos las celdas dentro del 30% de su valor nominal).

El patrón de supresiones también es calculado en base a una variable coste de pérdida de información. En nuestro caso la variable coste será la misma que la representada en las celdas (*gastot*) esto supone que a mayor valor de la variable dentro de la celda menor es la probabilidad de que ésta sea suprimida y por lo tanto menor la pérdida de información. La siguiente tabla refleja el patrón de supresiones óptimo según Argus:

Gastos financieros totales (gastot)

Tamaño act18	<= 100 trabajadores	>100 trabajadores	Total
Agropecuario y Pesca	**s**	.	**s**
Química	593301	1425519	2018820
Caucho plástico	388895	486771	875666
Metal	834529	2962455	3796984
Maquinaria	2274677	2740140	5014817
Materiales y manufacturas	4032117	8563599	12595716
Energía y construcción	**s**	**s**	557085
Otras actividades	**s**	**s**	18262444
Otros servicios	**s**	.	**s**
Total	25060426	18255122	43315548

7. Se observa que Argus ha añadido 3 supresiones complementarias a las que ya existían. Esta tabla será la publicada y se considerará segura para los límites de seguridad impuestos.

Existen múltiples formas de asegurar una tabla, tantas como recodificaciones posibles existen o especificaciones distintas hagamos (criterios de sensitividad, límites de seguridad...). La flexibilidad de Argus va a permitir crear tablas seguras tanto para publicaciones estándar como a la medida de un usuario concreto, sin que los datos pierdan utilidad estadística y capacidad de información.

Capítulo

5

Conclusiones y Futuro

Con la elaboración de este cuaderno se ha pretendido dar un repaso a las últimas técnicas desarrolladas en el ámbito de la Seguridad y Protección de datos estadísticos.(*) Además, hemos centrado la atención en el aspecto puramente estadístico de la confidencialidad, comprobando que existen y se están desarrollando métodos basados en los propios datos, que permiten dar la máxima calidad de información de la forma más segura.

Desarrollo de las Técnicas

Los avances en este campo se suceden a una gran velocidad y en el momento de la redacción de este cuaderno, surgen ya nuevas técnicas como mejora y alternativa a las expuestas. Ciertos aspectos comunes que deben contemplar estas nuevas técnicas, tanto para su aplicación en ficheros de registros como en tablas y en bases de datos, son los siguientes:

- Englobar a todas las fases de la producción de datos estadísticos (recogida, análisis y difusión)

- Evitar la sobreprotección, es decir, prescindir de supresiones innecesarias o redundantes en tablas, o de recodificaciones severas en ficheros de datos que lleven a una excesiva pérdida de información.

- Estudiar modelos probabilísticos que aporten medidas cuantitativas del riesgo de difusión, de forma que permitan la comparativa entre ficheros o tablas seguros y ayuden a elegir la mejor opción según el caso y las necesidades del usuario.

Software y Procedimientos Informáticos

No podemos obviar el desarrollo en paralelo de un software estandarizado que recoja la implementación de todas estas técnicas y que, a su vez, sea compatible con los nuevos sistemas de información y redes internacionales.

Nuevas versiones del software ARGUS presentado aquí, se están desarrollando como parte de un nuevo proyecto europeo para el control de la difusión de datos estadísticos. Una de las mejoras que se estudia incluir es la aplicación de microagregación para varias variables en el módulo μ-Argus, y el trabajo con grupos de tablas relacionadas en el módulo τ-Argus. También se estudia la compatibilidad con entradas y salidas para bases de datos relacionales (Acces, Oracle,....).

(*) La mayor parte de los métodos explicados fueron expuestos en la Conferencia sobre Protección y Seguridad de datos estadísticos llevada a cabo en Lisboa en Marzo de 1998.

No sólo a nivel europeo, sino mundial, existen paquetes de software epecializado que recogen los últimos adelantos en lo que a técnicas se refiere y proporcionan una cada vez mejor eficacia en tiempos de ejecución [11].

Bibliografía

[1] Baeyens,Y., Defays, D. (1998) "Estimation of variance loss following microaggregation by the individual ranking method". Proceedings of Statistical Data Protection 98' Lisbon.

[2] Cox,L.H., Zayatz, L..V. (1995) "An Agenda for Research in Statistical Disclosure Limitation". Journal of Official Statistics, Vol. 11, No.2, pp.205-220.

[3] Cox,L.H. (1998) "Some Remarks on Research Directions in Statistical Data Protection". Proceedings of Statistical Data Protection 98' Lisbon.

[4] Domingo-Ferrer,J. Mateo-Sanz,J.M. (1998) "A method for data-oriented multivariate microaggregation". Proceedings of Statistical Data Protection 98' Lisbon.

[5] Domingo-Ferrer,J., Mateo-Sanz,J.M., Sánchez del Castillo, R.X. (1999) "Cryptographic Techniques in Statistical Data Protection". Joint ECE/Eurostat Work Session on Statistical Data Confidentiality. Thessaloniki, Greece 99'.

[6] Domingo-Ferrer,J., Sánchez del Castillo, R.X., Castilla,J. (1998) "Dike: A Prototype for Secure Delegation of Statistical Data". Proceedings of Statistical Data Protection 98' Lisbon.

[7] Fienberg, S.E. (1994) "Conflicts Between the Needs for Access to Statistical Information and Demands for Confidentiality". Journal of Official Statistics, Vol. 10, No.2.

[8] Fischetti,M., Salazar,J.J. (1998) "Modelling and Solving the Cell Suppression Problem for Linearly-Constrained Tabular Data". Proceedings of Statistical Data Protection 98' Lisbon.

[9] Franconi, L.. (1999) "Level of safety in microdata: Comparisons between different definitions of Disclosure Risk and Estimation Models". Joint ECE/Eurostat Work Session on Statistical Data Confidentiality. Thessaloniki, Greece 99'.

[10] Garín, A., Ripoll, E. (1999) "Performance of μ-Argus in Disclosure Control of Uniqueness in Populations". Joint ECE/Eurostat Work Session on Statistical Data Confidentiality. Thessaloniki, Greece 99'.

[11] Giessing, S. (1998) "Looking for Efficient Automated Secondary Cell Suppression Systems: A Software Comparison". Proceedings of Statistical Data Protection 98' Lisbon.

[12] Gopal,R., Goes,P. (1998) "Confidentiality via Camouflage:The CVC approach to Database Query Management". Proceedings of Statistical Data Protection 98' Lisbon.

[13] Holvast, J. (1999) "Statistical Confidentiality at the European Level". Joint ECE/Eurostat Work Session on Statistical Data Confidentiality. Thessaloniki, Greece 99'.

[14] Horn,S., Morton, R., (1998) "Protecting Output Databases". Proceedings of Statistical Data Protection 98' Lisbon.

[15] Hundepool,A . J., Willenborg,L. (1998) "ARGUS for Statistical Disclosure Control". Proceedings of Statistical Data Protection 98' Lisbon.

[16] Hundepool,A . J., Willenborg,L. (1999) "ARGUS: Software from de Statistical Disclosure Control Project". Joint ECE/Eurostat Work Session on Statistical Data Confidentiality. Thessaloniki, Greece 99'.

[17] Jabine, T.B. (1993) "Statistical Disclosure Limitation Practices". Journal of Official Statistics, Vol. 9, No.2, pp.436-454.

[18] Keller-McNulty,S., Unger, E.A . (1993) "Database Systems:Inferential Security". Journal of Official Statistics, Vol. 9, No.2, pp.475-499

[19] Lambert, D. (1993) "Measures of Disclosure Risk and Harm". Journal of Official Statistics, Vol. 9, No.2, pp.313-331.

[20] Malvestuto, F.M., Moscarini, M. (1998) "An Audit Expert for Large Statistical Databases". Proceedings of Statistical Data Protection 98' Lisbon.

[21] McLeod, K.., George.J., Rae, A., Butler,R. (1998) "Investigating Key Qualities of an Automated Cell Suppression System". Proceedings of Statistical Data Protection 98' Lisbon.

[22] μ-ARGUS.Version 3.0. User´s Manual. Contributors: Hundepool, A.J., Willenborg,L.., Wessels, A., van Gemerden, L.., Tiourine, S., Hurkens,C.

[23] τ-ARGUS. Version 2.0. User´s Manual. Contributors: Hundepool, A.J., Willenborg,L.., Wessels, A., van Gemerden, L.., Fischetti, M., Salazar,J.J., Caprara, A.

[24] Ley de la Función Estadística Pública (9 de Mayo de 1989) Tit.1 Cap.III "Del secreto estadístico".