Estadística Española • Volumen 56, número 185 / 2014, pp. 279-296

Un sistema de cruce de información sensible sin compromiso de la privacidad de los datos fuente

David Pérez Fernández
Secretaría de Estado de Telecomunicaciones
y para la Sociedad de la Información

Resumen

El principal problema para la elaboración de tablas y cubos multidimensionales que crucen datos no públicos de distintas fuentes de información se encuentra en que requieren la firma de convenios para el intercambio de datos entre los órganos participantes o la cesión por parte de uno de los organismos de información confidencial. Ello supone un fuerte obstáculo para el desarrollo de estudios que impliquen múltiples fuentes de datos sensibles.

En el presente artículo se propone un mecanismo seguro que permite el cruce de datos entre dos o más organismos sin que ninguno de ellos deba ceder o comprometer la privacidad de aquellos datos de los que es responsable legal. El sistema requiere de un órgano intermediario que realiza el cruce de datos. El sistema garantiza, a su vez, la imposibilidad de acceso o almacenamiento de los datos tratados por parte del órgano intermediario. Este sistema de cruce de información hace uso de algoritmos criptográficos de clave pública-privada, clave simétrica y funciones de cegado.

Se plantean algunas generalizaciones del método propuesto. Por una parte, se amplía el método para permitir el cruce seguro de información entre más de dos entidades. Por otra parte, el sistema inicialmente propuesto, que permite contabilizar el número de registros para cada combinación de valores de las variables de clasificación, se generaliza para que pueda totalizar cualquier variable de explotación numérica, no sólo el conteo del número de registros cruzados.

En lo relativo al tratamiento de la confidencialidad, se incluye un método de chequeo del cumplimiento de los compromisos de confidencialidad estadística. Se propone un mecanismo de control de la confidencialidad aplicado al presente sistema de cruce de información sin que ninguno de los actores pueda deducir información confidencial sin ser previamente tratada.

Palabras clave: cruce de información, confidencialidad estadística,

Clasificación AMS: 62P99, 65K05, 62-07, 62Q05

A matching system for confidential information without source data privacy lost

Abstract

The main problem for the development of tables or multidimensional cubes crossing non public sensitive data from different information sources is signing confidentiality agreements. Usually one agent transfer sensible information to another agent. In this article a safe mechanism for crossing data between two or more agents without having to compromise data privacy is propossed. The system requires an intermediary agent that cross data. The system makes impossible to access or store data processed by the intermediary agent.

Proposed system makes use of cryptographic algorithms based on public-private and symmetric cryptography. Some generalizations of the method are also discussed. The method is extended to allow safe crossing of information between more than two agents. Moreover, the proposed system enables to create tables with number of records for each combination of values of the classification variables and also it can sum any numerical variable. Treatment of statistical confidentiality is also taken into account. A confidentiality control mechanism is outlined where no agent can deduce any confidential information.

Keywords: cross data, matching, statistical confidentiality,

AMS Classification: 62P99, 65K05, 62-07, 62 Q 05

1. Introducción

Hoy en día existe una enorme cantidad de información que proviene de distintas fuentes. El carácter horizontal propio de la función estadística hace del cruce de datos entre diferentes organizaciones, públicas o privadas, una necesidad para la realización de estudios estadísticos interdisciplinares. Estos estudios amplían el espectro de uso de la información original generalmente vinculada a la gestión dentro del negocio o la función propia de cada ente.

El principal problema para la elaboración de dichos estudios, compuestos de tablas y cubos multidimensionales, que fusione datos sensibles de distintas organizaciones, se encuentra en el requisito de la firma de convenios de colaboración para el intercambio de datos o la cesión por parte de alguno de los organismos de información confidencial sobre la que tiene la responsabilidad de su custodia, generalmente por imperativo legal.

Dado que el control de acceso a los datos no puede quedar totalmente garantizado cuando la información deja de estar en poder de una organización es usual que la organización de

mayor peso específico, o cuyos datos tienen mayor sensibilidad, en el sentido establecido por la Ley de Protección de Datos, sea la encargada de realizar el cruce de los mismos. Organizaciones que actúan como terceras partes en un estudio estadístico, por ejemplo Universidades o Institutos Estadísticos, en muchas ocasiones se ven privadas de la capacidad de realizar análisis de información sensible que proviene de distintas fuentes.

Este problema hace que la calidad, cantidad y rapidez del desarrollo de estudios estadísticos interdisciplinares sea muy inferior a la brindada por el enorme volumen de información disponible en la actualidad junto con la creciente capacidad de tratamiento de información.

En el presente artículo se propone un mecanismo seguro que permite el cruce de datos entre dos o más organismos sin que ninguno de ellos deba ceder sus datos confidenciales. El sistema requiere de un órgano intermediario que realiza el cruce de datos. El sistema garantiza la imposibilidad de acceso o almacenamiento de los datos tratados por parte del órgano intermediario.

Este sistema hace uso de algoritmos criptográficos de clave pública-privada y de funciones criptográficas de cegado. Las ideas planteadas son claras herederas de los sistemas de voto electrónico y de pago bancario anonimizado.

En el capítulo segundo se introduce un ejemplo previo para ilustrar las ideas del mecanismo general de cruce de datos que se propone en el siguiente capítulo. Se trata del cruce de datos ficticio de rentas, que provienen de la AEAT, con datos sobre los puntos de conducción, suministrados por la DGT.

En el capítulo tercero se propone el sistema general de intercambio de información sin compromiso de la privacidad de los datos fuente. Se incluyen distintas ampliaciones del mecanismo general.

Por una parte, se plantean generalizaciones para permitir el cruce seguro de información entre más de dos entidades. Por otra, se propone también una generalización para que pueda totalizarse cualquier variable de explotación numérica, no sólo el número de registros.

El cuarto capítulo está dedicado a la integración del tratamiento de la confidencialidad en el sistema de cruce de datos propuesto. Se incluye un método de chequeo del cumplimiento de los compromisos de confidencialidad estadística de forma previa a la publicación de la información. Se evita la posibilidad de conocer datos sensibles por ninguno de los órganos participantes. Se propone un mecanismo de control de la confidencialidad aplicado al presente sistema de cruce de información.

2. Un ejemplo previo

A continuación se presenta un ejemplo sencillo de cruce de datos sensibles entre la Agencia Española de Administración Tributaria, AEAT (organismo A) y la Dirección General de Tráfico, DGT (organismo D). En el ejemplo, se pretende realizar una tabla que cruce los datos referentes a puntos de conducción perdidos, de los que dispone la DGT, con el tramo de renta de los conductores, que posee la AEAT. Existirá también un organismo de intermediación,

que llamaremos organismo E. Como podrá comprobarse, dicho organismo no tendrá posibilidad de comprometer la privacidad de los datos suministrados.

Para elaborar la tabla citada es necesario cruzar los datos por un campo de identificación de cada registro único. En nuestro caso suponemos que la identificación se realiza a partir del Documento Nacional de Identificación, DNI de cada individuo. A continuación se describen los pasos del procedimiento de cruce de datos.

1. Agrupación en tramos de los datos

   La AEAT calcula el tramo de renta (TR) de cada ciudadano y la DGT elabora el listado de los puntos disponibles de cada conductor (PT). Ambas fuentes de datos deben poseer un campo identificador de cada individuo, en nuestro caso el DNI.

   Nota 1. Para fortalecer la seguridad del cruce de información, frente al órgano intermediador, es posible que los organismos participantes en el cruce de datos intercambien una clave simétrica para ocultar la variable de cruce, en nuestro caso el DNI.

2. Cegado y cifrado de las variables de clasificación

   Se aplica la función de cegado (C_A, C_D). Esta función y su inversa (que es aconsejable varíe en cada operación de cruce) sólo es conocida por cada entidad proveedora de información, por esta razón se emplea el sufijo correspondiente A, D.

   En nuestro ejemplo se propone la siguiente función de cegado sencilla:

   Imaginemos que los puntos o los tramos de renta se pueden tabular en n valores, que requieren m bits para codificarse. Si añadimos a esos m primeros bits una cierta cantidad de bits, de contenido aleatorio, hasta llegar a una longitud dada n^0, podemos aumentar el número de valores de cada clasificación y asignarlos de forma pseudoaleatoria. Dos valores idénticos de la variable de clasificación correspondiente a distintos registros pueden ser asignados a distintos valores. Es común incluir una operación de permutación adicional para dificultar el descifrado de los valores de las variables de clasificación.

   Después, el valor cegado se codifica con la clave pública del organismo emisor de la información para que sólo éste pueda obtener el valor original.

   Nota 2. Los objetivos que debe cumplir la función de cegado son:

   (a) Ocultar los valores de las variables de estudio. En nuestro caso esto se realiza por medio del completado aleatorio y la permutación del valor.

   (b) Ampliar la cantidad de valores asignados. Para que no sean fácilmente deducibles a partir de los valores de cruce.

   (c) Ser invertible, únicamente por la entidad que lo ha codificado.

3. Enlazado de datos y acumulación

   Los datos son enlazados por el organismo neutro (órgano E). Previamente se ha descifrado el código de identificación del registro empleando la clave privada del

organismo E. Con ello se garantiza que nadie pueda interceptar el fichero y obtener las identificaciones originales.

Después de realizar el enlazado se acumulan los registros que tengan idénticos valores de las variables cegadas C_A(TR) y C_A(PT). Se genera una nueva columna que recuenta las repeticiones de las combinaciones de estos valores. Este conteo no devolverá los valores definitivos puesto que la función de cegado, C, ha ampliado el número original de valores de las clasificaciones TR y PT.

Nota 3. En este ejemplo se realiza una tabla que cuenta el número de ciudadanos que para cada tramo de renta ha perdido una determinada cantidad de puntos. Es posible realizar la suma de otras variables numéricas, no el simple conteo del número de registros, pero requiere un mecanismo más complejo de cruce que se explicará más adelante.

4. Inversión del cegado y agregación por el informante, AEAT

   La nueva tabla agregada es enviada al organismo no peticionario para que aplique la inversa de la función de cegado, C_A^{-1}.

   Se realiza una nueva acumulación de registros puesto que la función de cegado ha asignado valores distintos a un mismo valor de la variable de estudio, TR, en diferentes registros.

5. Inversión del cegado y acumulación por el peticionario, DGT

   El órgano peticionario recibe la información y vuelve a realizar el proceso anterior, inversión del cegado, aplicando C_D^{-1}, y acumulación. De este modo se obtienen los valores originales de la variable PT.

   Puede requerirse una nueva acumulación de valores puesto que la función de cegado puede haber asignado varios valores distintos a un mismo tramo de la variable de estudio, PT, para dos registros distintos.

Finalmente se ha obtenido el fichero deseado que contiene el número de individuos que se encuentran en cada tramo de renta y puntos de conducción que provienen de la DGT.

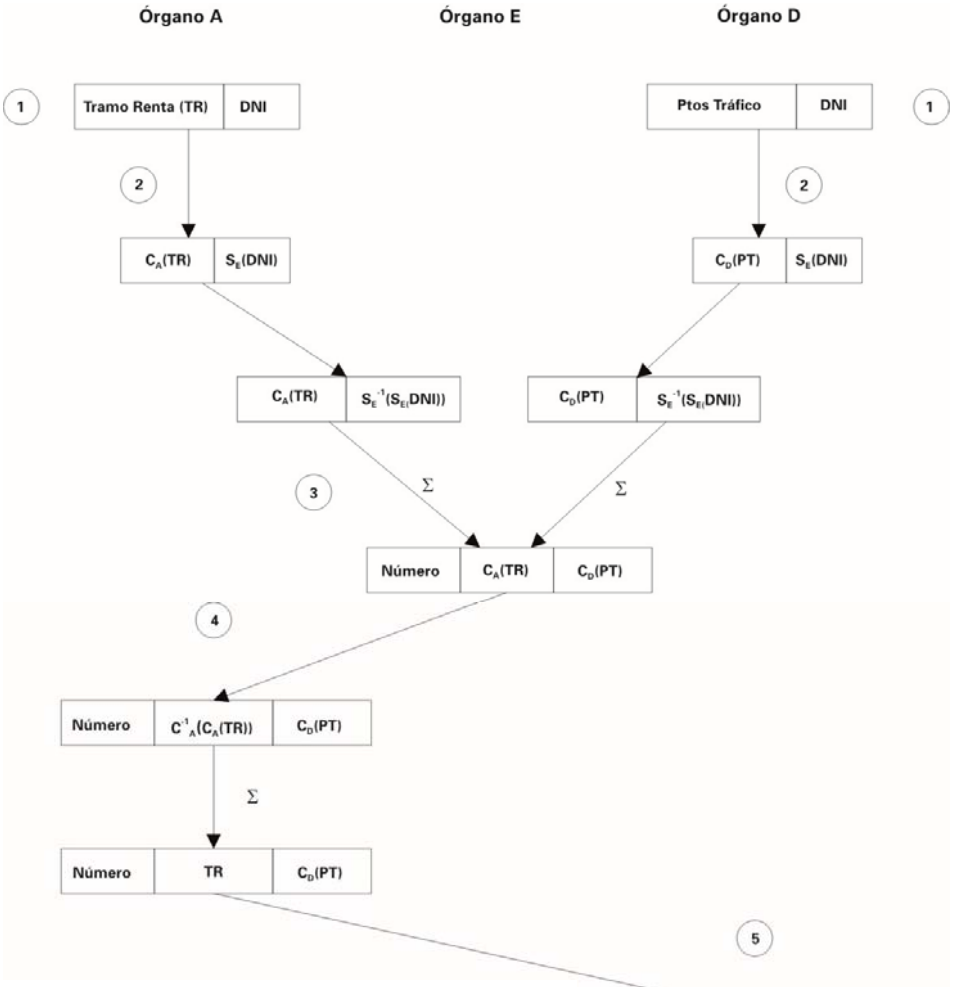
Se ha realizado el cruce de datos sensibles sin que se haya requerido la cesión de los datos de ninguno de los dos organismos. El organismo intermediador tampoco ha podido deducir información alguna sobre los valores cruzados.

Nota 4. Como se ha comentado anteriormente, es posible incrementar la seguridad del sistema intercambiando una clave simétrica que permita el cifrado de los valores de la columna empleada para el cruce e identificación de los registros. Esto permite asegurarse de que el órgano intermediador no puede tratar de deducir nada sobre la información sensible empleada en el cruce.

Nota 5. En el diagrama se ha incluido una clave adicional, S_E, perteneciente al órgano intermediario *E* para mejorar la seguridad en el intercambio de información. Es posible incluir una clave simétrica adicional conocida por los órganos *A* y *D* para la codificación de los identificadores de registro. Con ello el órgano intermediario desconocerá incluso los identificadores de los registros que cruza.

En el siguiente gráfico se muestran los pasos del procedimiento descrito:

Gráfico 1



C_A: función de cegado del órgano *A*.

C_D: función de cegado del órgano *D*.

C_A^{-1}: función de inversión de cegado del órgano *A*.

C_D^{-1}: función de inversión de cegado del órgano *D*.

S_E: clave pública órgano *E*.

S_E^{-1}: clave privada órgano *E*.

Σ: acumulación de valores.

3. Planteamiento y generalizaciones

Tenemos dos órganos, *A* y *B*, que quieren cruzar datos cuyas variables de estudio contienen información sensible. El organismo *A* actúa como peticionario de la información y el organismo *B* como el segundo informante. Se designa como órgano *E* al ente intermediador que realiza el cruce de datos cegados.

Nota 6. El sistema propuesto es independiente de la designación de peticionarios e informantes. Sólo han sido designados a título informativo para trazar un orden temporal en el proceso de intercambio de información.

Los pasos generales del procedimiento seguro de cruce de datos son:

1. Intercambio de clave simétrica
   Paso opcional, {Órganos *A*,*B*}: Comunicación entre los órganos *A* y *B* del mecanismo invertible de ocultación de valores de la variable de cruce (por ejemplo, una permutación de los mismos por una clave simétrica).

2. Cegado y cifrado de identificadores

   {Órganos A,B}: Cegado de las variables de estudio y cifrado con la clave pública del órgano intermediador E (para evitar su interceptación por terceros del envío de información).

3. Enlazado de datos y acumulación por el órgano neutro

   {Órgano E}: Cruce de datos por el organismo neutro E. Previamente se ha descifrado el código de identificación de cada registro (empleando la clave privada del órgano E).

   {Órgano E}: Posteriormente al cruce, se acumulan registros que tengan idénticos valores de las variables de estudio formando una primera tabla, a la que llamaremos tabla 0, con las variables de estudio cegadas. Se genera una nueva columna que recuenta las repeticiones de registros para dicha combinación de valores de las variables de clasificación.

   Nota 7. El cruce de datos se ha realizado a partir de los identificadores de registro que, como se ha dicho, han podido ser previamente cifrados con una clave simétrica para fortalecer la seguridad de la transmisión de la información y frente al órgano intermediador.

4. Inversión del cegado y agregación por el informante

   {Órgano B}: La nueva tabla agregada, tabla 0, es enviada al organismo no peticionario, B, para que aplique la inversa de la función de cegado a las variables que previamente ha cegado.

   Se realiza una nueva acumulación de valores puesto que la función de cegado ha asignado varios valores distintos a un mismo valor de la variable de estudio original. De este modo, se crea una nueva tabla, a la que llamaremos tabla 1, con un mayor grado de agregación que la anterior tabla 0.

5. Inversión del cegado y agregación por el peticionario

   {Órgano A}: El órgano peticionario recibe la nueva tabla, tabla 1, y realiza el proceso de inversión del cegado y acumulación de registros. Se forma la tabla final con los valores originales de todas las variables de estudio.

Finalmente se obtiene la tabla con el cruce de información deseado sin se haya requerido la cesión de datos sensibles por parte de ninguno de los organismos. Tampoco el ente intermediador ha podido analizar los valores cruzados puesto que han sido previamente cegados, y de forma opcional se han cifrado los identificadores, haciendo posible su cruce pero no la identificación del registro.

Nota 8. Aportación de múltiples variables de estudio por cada organismo

En el modelo propuesto se ha incluido una única variable de estudio aportada por cada organismo. Es sencillo ver cómo cada organismo puede asignar valores a los registros que representen valores de múltiples variables de clasificación. El cruce se realiza del mismo modo pero cuando el organismo realiza la tarea de inversión de la función de cegado enlaza los valores de las múltiples variables de clasificación con sus valores correspondientes.

3.1 Cruce de datos entre más de dos organismos

En el primer apartado de esta sección se ha expuesto el mecanismo general para el cruce seguro de información entre dos organismos sin cesión de datos, este esquema es fácilmente generalizable a cualquier número de organismos participantes en el cruce. Es decir, es posible formar tablas multidimensionales, o cubos de datos, que contengan información sobre distintas variables de clasificación aportadas por múltiples organismos.

Siguiendo nuestro primer ejemplo, se podría incluir información sobre la situación laboral del sujeto provista por un tercer organismo, por ejemplo el INSS.

Para poder generalizar el mecanismo de cruce información hay que tener en cuenta:

- En caso de que se realice el paso opcional de intercambio de clave simétrica, *S*, para la encriptación de los identificadores de registro, los organismos deben establecer un mecanismo eficaz de intercambio de dichas claves. A partir de las claves públicas de los organismos es fácil intercambiar la clave *S* sin que pueda ser interceptada por terceros ni encontrarse disponible para el organismo intermediador.
- En el mecanismo inicialmente descrito, la tabla inicial es acumulada por el organismo intermediador *E*, enviada al órgano *B* y posteriormente llega hasta el órgano peticionario *A*. Es decir, el recorrido es: $E \underset{T_0}{\rightarrow} B \underset{T_1}{\rightarrow} A$ Se recuerda que la tabla intercambiada en cada paso no es la misma puesto que van agregándose los registros y cada órgano participante va invirtiendo las funciones de cegado sobre los valores de las variables que ha aportado.

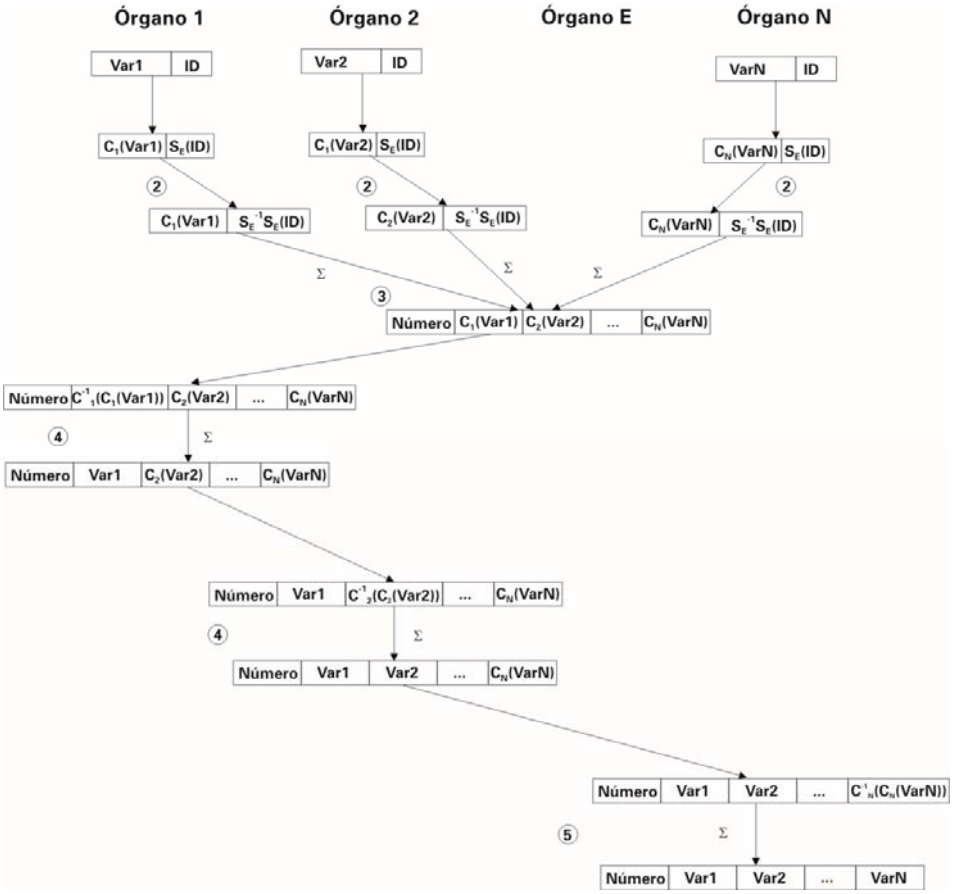
En caso de que existan más de dos organismos, que llamaremos O_i, este recorrido debe ser más amplio pero funcionalmente equivalente:

$$E \underset{T_o}{\rightarrow} O_1 \ldots \underset{T_{n-1}}{\longrightarrow} O_n \qquad [3.1]$$

El único requisito es que en esta secuencia el último organismo, O_n, sea el órgano peticionario de la información.

En el siguiente gráfico se muestran los pasos del procedimiento general:

Gráfico 2



3.2 Otras variables de explotación numéricas distintas del número de registros

En el procedimiento descrito anteriormente se realiza el conteo de registros que tienen unos valores determinados de las variables de clasificación del estudio. Es posible realizar la suma de otras variables cuantitativas.

A continuación se explica con un ejemplo la suma de una variable de explotación numérica proporcionada por dos organismos. Ambos órganos desean sumar, para unos tramos dados de una variable de estudio, una determinada cantidad que poseen ambos organismos sin ceder sus datos ni desvelar su privacidad.

Supongamos que una empresa puede deber dinero a dos organismos (A y B). Uno de los organismos, o un tercer órgano, desea realizar el estudio de las deudas contraídas con los órganos A y B según el número de empleados de la empresa. El número de empleados actúa como la variable de clasificación del estudio, una vez agrupada en tramos. El objetivo del estudio es llegar a obtener una tabla que muestre, para ambas administraciones colaboradoras, la cantidad total adeudada por las empresas por tramos de número de empleados.

Para que pueda realizarse el cruce de datos, sin que el órgano de intermediación obtenga los valores de las deudas contraídas con los órganos participantes a todas las empresas españolas, se establece el siguiente mecanismo de intercambio de datos:

- Se realiza una estimación del valor máximo de las sumas posibles de la variable cuantitativa (importe de las deudas) para todos los valores de las variables de estudio (tramo de número de empleados). Se busca un número primo, p, superior a dicho valor máximo y suficientemente alto para que su inversa no pueda obtenerse de forma sencilla.

- Multiplicando por un valor inferior a dicho primo, e, y realizando el módulo por ese número primo p se produce una permutación de todos los valores inferiores. Esta permutación es invertible por medio del producto por un valor único d, llamado inverso de e, que cumple $d \cdot e = 1 \ mod(p)$.

  Nota 9. Esta propiedad se cumple por haber elegido un número primo p y trabajar en el cuerpo algebraico $\mathbf{Z}_p$.

  Nota 10. En caso de tratarse de valores positivos, si no queremos que el valor nulo sea enviado por esta permutación al mismo valor debemos desplazar todos los valores una cierta cantidad. En caso de tratarse de valores negativos debemos hacer una estimación de valor mínimo y aplicar una traslación suficiente de los valores para trabajar en el cuerpo $\mathbf{Z}_p$.

- El valor permutado no puede ser interpretado por el organismo de intermediación, sin embargo, puede sumarlo para el mismo valor de las variables de clasificación del estudio (en nuestro caso, el tramo de número de empleados de cada registro). Después de haberse realizado el cruce de datos y la posterior acumulación, ambos órganos pueden multiplicar por el factor de inversa, d, y recuperar la suma original dado que se cumple:

. $$d \cdot (e \cdot v_1 + e \cdot v_2) = v_1 + v_2 \quad [3.2]$$

siendo v_i valores inferiores a p. Esta propiedad se cumple por haber elegido p primo y tener $\mathbf{Z}_p$ propiedades de cuerpo algebraico.

4. Control de la confidencialidad

4.1 Confidencialidad estadística

Existe una creciente demanda de información socio-económica y, por otra parte, la obligación legal y ética de proteger la privacidad de los individuos y las empresas que son la fuente de dichos datos estadísticos. El control de la confidencialidad implica la resolución del problema de maximizar la información proporcionada manteniendo las restricciones que impone la protección de los datos estadísticos confidenciales.

El conjunto de métodos que intentan dar solución a este problema se denominan técnicas de control de la confidencialidad estadística (Statistical Disclosure Control, SDC). En la actualidad existe una gran variedad de técnicas de control de la confidencialidad estadística, ver [1], [2] y [3]:

- *Métodos de restricción*: Limitación de la información publicada:
  - *Supresión de celdas (CSP)*: Se suprimen tanto las celdas que contienen información confidencial, llamadas supresiones primarias, como las que puedan servir para inferir las primeras, supresiones secundarias.
  - *Métodos de recodificación*: Agrupación de valores de variables de clasificación para ocultar los datos confidenciales de determinados colectivos minoritarios.
  - *Publicación de intervalos*: En lugar de publicar datos individuales, se muestran los márgenes del intervalo seguro en el que se encuentra el verdadero valor.
- *Métodos de perturbación*: Alteración de los datos originales proporcionando unos nuevos datos sintéticos protegidos:
  - *Ruido (CTA)*: Pequeñas variaciones en los valores o en las variables de clasificación de modo que desaparezcan o queden ocultas las celdas deseadas.
  - *Sustitución*: Intercambio de valores dentro de una variable de clasificación para su tratamiento estadístico sin posibilitar su identificación.
  - *Redondeo aleatorio (CRP)*: Redondeo aleatorio de los valores confidenciales para mantener su confidencialidad.
  - *Microagregación*: Sustitución de los valores confidenciales valores medios dentro del grupo al que pertenecen.

En las siguientes secciones de este apartado se estudiará de qué forma puede integrarse el control de la confidencialidad estadística con el sistema de cruce de datos propuesto.

En primer lugar, se estudiará cómo puede verificarse que la tabla que será publicada cumple con los criterios de confidencialidad previamente establecidos. Esto procedimiento debe realizarse sin que ninguno de los intervinientes tenga posibilidad de obtener información confidencial.

En segundo lugar, se propone un mecanismo de cruce de datos que incorpora control de la confidencialidad estadística por el sistema de restricción de información llamado supresión de celdas (CSP).

4.2 Control de cumplimiento de los criterios de confidencialidad

De forma previa a cada estudio los organismos deben emplear criterios de sentido común en la elección de los tramos de las variables de clasificación del estudio para minimizar la probabilidad de compromiso de la confidencialidad de los datos finales tabulados. Cada organismo debe ser consciente de que la tabla final objeto del estudio muestra datos con un nivel de desagregación superior a la agrupación en tramos original.

Por otra parte, sólo cuando obtenemos la tabla final podemos comprobar si se vulneran los límites de confidencialidad de los datos establecidos para el estudio. Se propone el siguiente mecanismo de control del cumplimiento de los objetivos de confidencialidad previo a la obtención de la tabla definitiva.

1. Cruce de los datos por el órgano intermediador

   Después de que el órgano intermediador ha realizado el cruce de datos de los registros, con los valores de las variables de estudio cegadas, se envía la tabla al órgano informante y al peticionario.

   Se realizan dos procesos similares de forma paralela. La intención de la duplicidad del proceso es dotar de mayores garantías sobre la permutación de valores realizada por los organismos intervinientes (debe mantenerse la misma permutación en ambos procesos):

2. Envío desde el órgano intermediario.

   El organismo intermediario envía la tabla cruzada con los valores de clasificación cegados al organismo peticionario (para la verificación por parte del órgano informante) y también al órgano informante (para la verificación final por parte del órgano peticionario).

3. Verificación por el órgano informante

   – Permutación por el órgano peticionario

     El órgano peticionario recibe la tabla cegada. Realiza una inversión del cegado y realiza una permutación de los valores de aquellas variables de clasificación que ha cegado.

   – Permutación por el órgano informante

     En segundo lugar, el órgano peticionario envía al órgano informante la tabla para que realice la misma operación de inversión del cegado y permutación de los valores de las variables de clasificación.

Finalmente, el órgano informante chequea los límites de confidencialidad y envía al órgano intermediador la tabla con todos los valores de las variables de clasificación permutados.

4. Verificación por el órgano peticionario

   De forma paralela, se realiza el mismo proceso pero en orden inverso.

   – Permutación por el órgano informante

     El órgano informante recibe la tabla cegada. Realiza una inversión del cegado y realiza una permutación de los valores de las variables de clasificación.

   – Permutación por el órgano peticionario

     Después, el órgano informante envía al órgano peticionario la tabla para que realice la misma operación de inversión del cegado y permutación de los valores de las variables de clasificación.

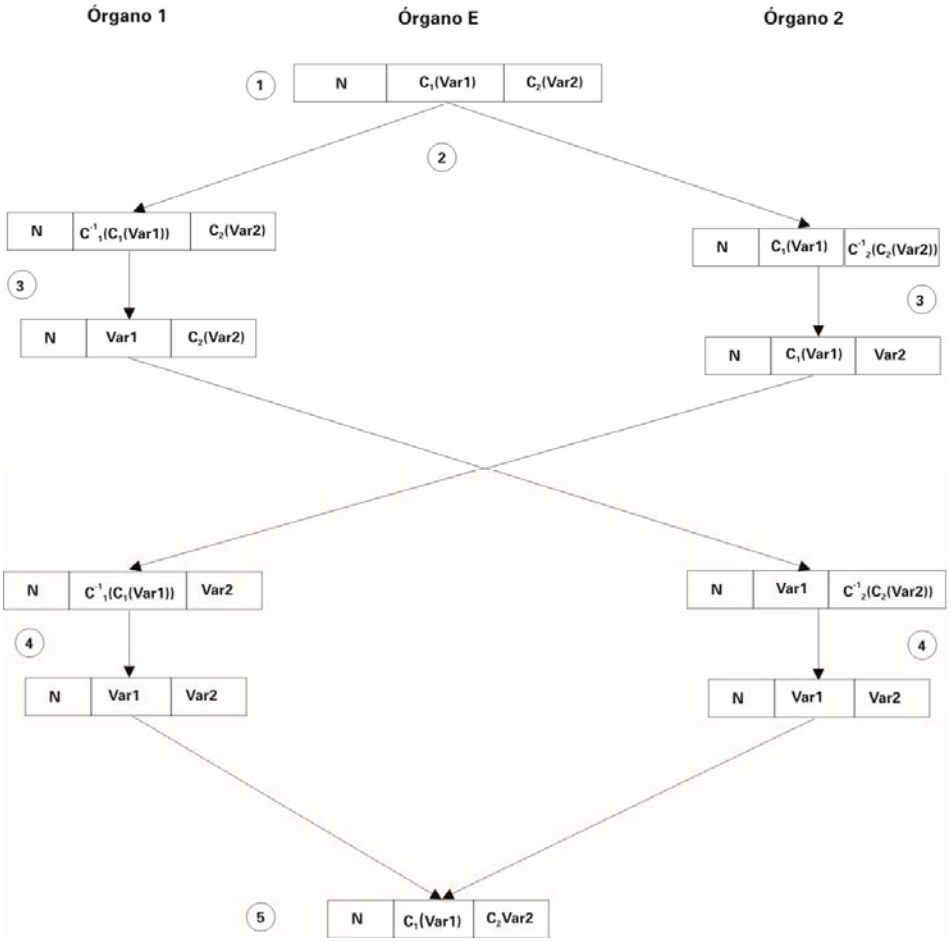
     Finalmente, el órgano peticionario chequea los límites de confidencialidad y envía al órgano intermediador la tabla con todos los valores de las variables de clasificación permutados.

5. Chequeo de compromiso de confidencialidad por el órgano intermediario

   El órgano intermediario chequea que ambas tablas coinciden y respetan los límites de confidencialidad. En caso afirmativo, se autoriza el cruce de datos siguiendo el procedimiento anteriormente descrito. En caso contrario, se notifica a las partes la denegación del cruce de datos.

   En el siguiente gráfico se muestran los pasos del procedimiento descrito:

Gráfico 3



En el caso de que no se respeten los límites de confidencialidad, caben varias posibilidades. Se puede proponer nuevos tramos para las variables de clasificación y chequear otra vez si se cumple el compromiso de confidencialidad o realizar algún tipo de control sobre la tabla que ha de publicarse. En el siguiente apartado se propone la aplicación de la metodología de supresión de celdas aplicada al método de cruce de datos propuesto.

4.3. Control de la confidencialidad

En este apartado se utiliza la metodología conocida como supresión de celdas consistente en la no publicación de las celdas confidenciales y aquellas que puedan servir para deducirlas. Es posible integrar otras metodologías en el sistema de cruce de datos que analizamos.

Como se ha dicho, si se empleara la agregación de tramos de las variables de clasificación bastaría con realizar el chequeo del cumplimiento de los criterios de confidencialidad y, en caso de ser necesario, reagrupar, de forma más grosera, los tramos de las variables de clasificación del estudio.

Cuando se emplea el método de supresión de celdas se ocultan aquellas celdas que no cumplen con los criterios de confidencialidad. Cuando las tablas publicadas incluyen marginales o subtotales en los que participan celdas suprimidas es posible deducir algunos de los valores ocultos. Los valores que se quieren ocultar en primera instancia se les llama *supresiones primarias* y aquellos que se ocultan para que no puedan deducirse las supresiones primarias, serán llamados *supresiones secundarias*. Para una exposición completa de esta metodología y su resolución óptima ver [4]. Una implementación abierta de los algoritmos expuestos en el artículo anterior puede verse en [5].

Veamos ahora cómo es posible modificar el procedimiento de cruce de información para incluir el control de la confidencialidad por medio de la metodología de supresión de celdas.

Se trata de realizar un proceso similar al inicialmente expuesto pero modificando el paso final. Antes, en el paso final, los órganos informante y peticionario realizaban el chequeo de conformidad con el cumplimiento de las condiciones de confidencialidad.

En esta ocasión aplicarán la metodología de supresión de celdas para buscar aquellas supresiones secundarias que garantizan el cumplimiento de los criterios de confidencialidad. Después del cálculo de la supresión de celdas óptima, realizado por ambos órganos participantes, han de ponerse de acuerdo en la solución óptima que emplearán (dado que ésta solución es óptima pero no necesariamente única).

El cálculo de las supresiones secundarias ha de realizarse sobre la tabla de celdas permutadas. Esta tabla contiene los mismos subtotales y marginales que la tabla final no permutada pero las celdas interiores no se corresponden con los valores de los tramos originales.

Las permutaciones de los valores de las variables de clasificación, realizadas en el proceso de chequeo, deben ser memorizadas por los órganos intervinientes. La razón es que dichos órganos deberán proporcionar los valores cegados originales que deben ser ocultados.

1. Cálculo de las supresiones secundarias

   A partir de las ecuaciones conocidas que relacionan celdas de la tabla objetivo se plantea el problema de supresión de celdas, conocido como problema CSP. Ambos órganos participantes emplean la misma metodología de resolución de problema CSP (por ejemplo por medio de la implementación abierta [6]) y se calculan las supresiones secundarias. Si existen múltiples soluciones óptimas se selecciona una de ellas, poniéndose de acuerdo los órganos intervinientes. El órgano intermediador puede chequear que las supresiones secundarias coinciden. Se inician dos procesos paralelos similares. Se describe sólo uno de ellos:

2. Inversión de la permutación y cegado de las supresiones secundarias por el órgano informante

El órgano informante invierte la permutación de los valores de las variables de clasificación de aquellas celdas que deben ser suprimidas y ciega las celdas para que no puedan ser descubiertas por el órgano peticionario.

Después envía al órgano peticionario la colección (ampliada al haber sido cegadas) de celdas que deben ser suprimidas.

3. Inversión de la permutación y cegado de las supresiones secundarias por el órgano peticionario

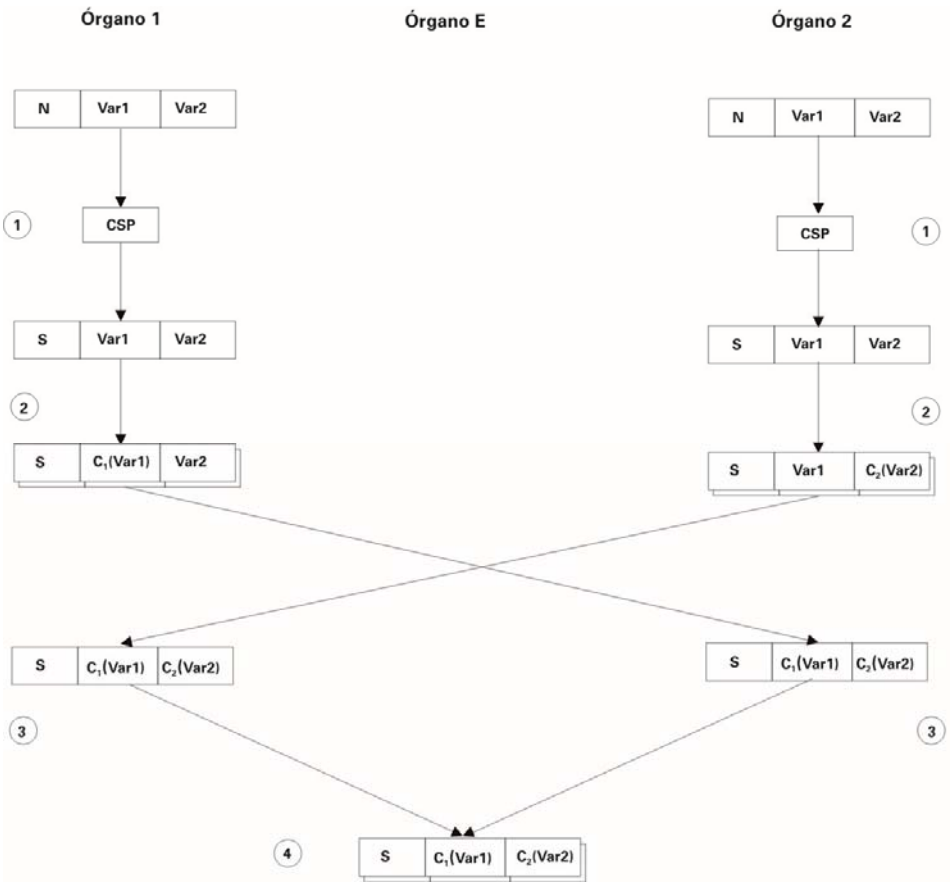
   El órgano peticionario realiza la misma operación de inversión de la permutación realizada para el chequeo de confidencialidad y posteriormente ciega los valores de las variables de clasificación de las celdas que deben ser suprimidas.

4. Control por el órgano de intermediación

   El órgano de intermediación recibe las colecciones ampliadas de celdas que deben ser suprimidas. Las elimina de la tabla cegada original e inicia el proceso de cruce de datos habitual.

En el siguiente gráfico se muestran los pasos del procedimiento descrito:

Gráfico 4



S: supresiones primarias y secundarias.

CSP: cell suppresion problem.

5. Conclusiones

En el presente artículo se ha expuesto un sistema de cruce de información sensible sin compromiso de la privacidad de los datos de origen. Se han incluido algunas generalizaciones como la participación de más de dos órganos de datos fuente o la suma de variables explotación general, no sólo el conteo de registros según sus valores de las variables de clasificación.

Se ha mostrado cómo es posible integrar el chequeo y control efectivo de la confidencialidad con el sistema de cruce de información propuesto. Se ha propuesto un ejemplo de integración con la metodología de supresión de celdas, CSP.

En el contexto de la Administración, este sistema puede ser de interés general para cualquier instituto estadístico u órgano encargado de la interoperabilidad de datos entre Administraciones. Cualquiera de ellos puede actuar como órgano de intermediación implementando informáticamente el sistema de cruce de datos y poniendo a disposición de terceros las herramientas necesarias para realizar el cruce de datos. La puesta a disposición pública de este sistema al resto de organizaciones o terceras partes interesadas (por ejemplo, universidades) permitirá realizar estudios estadísticos que relacionen distintas fuentes de datos sensibles.

Tanto en el contexto legal español, por medio de la *Ley 12/1989*, de 9 de mayo, de la Función Estadística Pública, [8], como el correspondiente mandato de Eurostat se avala la posición intermediadora de estos órganos en lo referente al cruce de datos estadísticos.

Por otra parte, la *Ley 11/2007* de Acceso electrónico de los ciudadanos a los servicios públicos, ver [9], promueve la cooperación entre Administraciones para el impulso de la administración electrónica, el intercambio de información y la transferencia de tecnología entre las mismas.

Finalmente, las iniciativas de europeas (directiva de reutilización de la Información del Sector Público) y nacionales (incluidas en el Esquema Nacional de Interoperabilidad) marcan el desarrollo estratégico del llamado Open Data y la reutilización de la información en el Sector Público (RISP), ver [7].

El sistema de cruce de datos propuesto es una herramienta destinada a facilitar la reutilización efectiva de datos de diversas fuentes por terceros organismos o empresas que no participen en su generación y gestión cotidiana pero que aporten valor adicional a los mismos.

En cuanto a la implementación efectiva del sistema de cruce de información puede realizarse por medio de un proyecto abierto compartido entre institutos estadísticos europeos o a nivel nacional bajo el paraguas del Instituto Nacional de Estadística o el Ministerio de Hacienda y Administraciones Públicas, actual impulsor del Esquema Nacional de Interoperabilidad.

Las principales funcionalidades del sistema podrían exponerse al resto de los organismos por medio de servicios web estándares ejecutados en la nube y todo el proceso puede controlarse por medio de un gestor de flujos de trabajo estándar.

Se han realizado pruebas sobre el rendimiento de las tareas básicas de cegado, cifrado y cruce sobre colecciones de veinte millones de registros. Las operaciones más lentas no han superado unos pocos minutos, en ordenadores de sobremesa con procesares Intel *i*5. Los tamaños de las tablas cegadas no han superado las 200 Mb.

Referencias

WILLEMBORG, DE WALLL (1996), «Statistical Disclosure Control in practice, Lecture notes in Statistics, Springer Verlag». New York.

DUNCAN. KELLER-MCNULTY, STOKES (2001), «Disclosure Risk vs Data Utility: the R-U Confidentiality Map», *Technical Report LA-UR-01-6428. Statistical Science Group,* Los Álamos Laboratory, NM.

TREWIN (2006), «Principles and guidelines of good practice for managing statistical confidentiality and microdata access», UNECE.

FISCHETTI, SALAZAR (2001), «Solving the cell suppression problem on tabular data with linear constraints, Management Science».

INSTITUTO NACIONAL DE ESTADÍSTICA. *Revista Estadística Española número 179*. Tercer cuatrimestre de 2012

JCSP 2.0
http://administracionelectronica.gob.es/ctt/verPestanaGeneral.htm? idIniciativa=322

PLANES RISP DEL SECTOR PÚBLICO ESTATAL
http://datos.gob.es/content/guia-de-aplicacion-de-norma-tecnica-de-interoperabi

LEY DE LA FUNCIÓN ESTADÍSTICA PÚBLICA
http://www.boe.es/buscar/doc.php?id=BOE-A-1989-10767

LEY DE ACCESO ELECTRÓNICO DE LOS CIUDADANOS A LOS SERVICIOS PÚBLICOS
https://www.boe.es/buscar/doc.php?id=BOE-A-2007-12352