Postprocesado para Microagregación Multivariante: un Estudio con Datos Reales

G. Pujol, A. Martínez-Ballesté, J. M. Mateo-Sanz y A. Solanas, *Member IEEE*

***Resumen*— La privacidad de datos es el equivalente en la Sociedad de la Información al derecho fundamental a la privacidad y a la vida privada, incluido por las Naciones Unidas en la Declaración Universal de los Derechos humanos de 1948.**

Con las nuevas Tecnologías de la Información y de las Comunicaciones (TICs), en la actualidad se pueden almacenar y analizar grandes cantidades de datos personales. De este modo, si no se toman las medidas de seguridad adecuadas, la privacidad individual puede estar en peligro.

Hasta la fecha ha sido propuesta una gran cantidad de técnicas de control de la revelación estadística, cuyo propósito es proteger la privacidad de los individuos, cuyos datos son almacenados. Entre éstas, el *k*-anonimato es una propiedad prometedora que, si se consigue adecuadamente, puede ayudar a este propósito.

En el presente artículo, proponemos un método de postprocesado que puede ser usado tras la aplicación de un algoritmo de *k*-anonimato, para reducir los errores resultantes de la agregación de datos. Los resultados experimentales demuestran que el método propuesto es útil tanto en datos reales como sintéticos.

***Palabras clave*— Análisis multivariante *(multivariate analysis)*, clusterización *(clustering)*, *k*-partición *(k-partition)*, *k*-Anonimato *(k-anonymity)*, microagregación *(micro-aggregation)*, privacidad *(privacy)*.**

I. Introducción

LAS Tecnologías de la Información y de las Comunicaciones (TICs) están presentes en muchos aspectos de nuestra vida diaria. Su uso masivo por parte del gran público conlleva la reunión de grandes cantidades de información personal. Existen muchas situaciones de la vida real en las que se almacena información personal:

- *E-comercio*. El comercio electrónico implica la recogida automática de enormes cantidades de información del consumidor. Estos datos, que recogen diversas compañías, se comparten con subsidiarios y socios.
- *Información sanitaria*. La sanidad es un sector muy sensible con restricciones muy estrictas es cuanto al manejo de datos personales se refiere [1][2].
- *Servicios basados en la localización*. Los servicios relacionados con la posición actual del usuario que se utilizan en los teléfonos móviles crecen rápidamente. Si las consultas que realiza un usuario a un servidor basado en la localización no son manipuladas con seguridad, podría ser posible deducir los hábitos de consumo del usuario [3].
- *Tecnología RFID*. El despliegue masivo de la tecnología de IDentificación por Radio Frecuencia (RFID) es una realidad. Por un lado, esta tecnología incrementará la eficiencia de las cadenas de suministros y muy pronto puede llegar a reemplazar los códigos de barras. Por otro lado, la existencia de etiquetas RFID en prácticamente todos los objetos de uso cotidiano puede verse como un problema de privacidad [4].

Además de los puntos expuestos hasta ahora. Las agencias de estadística son el caballo de batalla de la protección de la privacidad. La mayoría de los países disponen de legislación referente a las agencias nacionales de estadística para garantizar confidencialidad estadística cuando éstas publican datos recogidos de ciudadanos o compañías; ver [5] para la legislación de la Unión Europea, [6] para Canadá y [7] para USA. Proteger la privacidad individual es una cuestión clave no sólo para agencias de estadística, sino para muchas instituciones y empresas, las cuales deben desarrollar y aplicar técnicas para garantizar el mayor grado posible de privacidad personal.

Si nos centramos en la protección de los datos publicados por estas agencias de estadística, una posible solución es la agregación de los datos originales antes de su publicación. esta idea se inició en campo del Control de la Revelación Estadística (*Statistical Disclosure Control*, SDC). Agregando los datos, es estadísticamente improbable relacionar los datos publicados con un individuo y, al mismo tiempo, se preserva la utilidad de los datos. Agregando los datos originales, podemos obtener registros *k*-anónimos que pueden ser publicados preservando un grado de privacidad (*i.e.* para cada registro dado, hay otros *k*-1 registros indistinguibles de modo que no es posible re-identificar individuos inequívocamente). Sin embargo, esta solución no es sencilla porque los elementos que componen los datos deben ser correctamente agrupados antes del proceso de agregación. De los contrario, la pérdida de información asociada a la agregación podría dejar inservibles los datos agregados. La agrupación de datos con

A. Solanas, G. Pujol y A. Martínez-Ballesté pertenecen al Departamento de Ingeniería y Matemáticas de la Universidad Rovira y Virgili, Av. Països Catalans 26, 43007 Tarragona (correo e.: agusti.solanas@urv.cat; antoni.martinez@urv.cat; gloria.pujol@urv.cat), y al grupo de investigación CRISES (http://crises-deim.urv.cat/webCrises).

J. M. Mateo-Sanz pertenece al Área de Estadística de la Universidad Rovira y Virgili, Av. Països Catalans 26, 43007 Tarragona (correo e.: josepmaria.mateo@urv.cat).

restricciones en el tamaño de los clusters es un problema NP cuando los datos son multivariantes [8]. Por ello es necesario utilizar heurísticas para agregar correctamente los datos multivariantes.

A. Contribución y plan del artículo

En este artículo proponemos un método de postprocesado para reducir la pérdida de información asociada a procesos de protección de la privacidad mediante microagregación. Utilizando nuestro método, los clusters generados son más compactos y el error debido al proceso de agregación-anonimización es menor.

El resto del artículo está organizado como sigue: en la Sección II se da un pequeño resumen de los conceptos básicos de SDC. Después, en la Sección III, se detalla nuestra propuesta, y en la Sección IV se proporcionan algunos resultados experimentales. Finalmente, el artículo concluye en la Sección V.

II. ANTECEDENTES

En esta sección proporcionamos al lector algunas ideas básicas sobre el Control de la Revelación Estadística y algunas de sus técnicas (como la microagregación y el *k*-anonimato).

A. Control de la revelación Estadística

El Control de la Revelación Estadística (SDC) busca transformar los datos de una manera que puedan ser publicados al mismo tiempo que se preserva su utilidad y la confidencialidad estadística, significando esto último evitar la revelación de información que pueda relacionarse con un individuo o empresa específico que participó en la recogida de información. El SDC puede aplicarse a datos en diferentes formatos, lo que induce la siguiente clasificación de las técnicas SDC:

- **Protección de datos tabulares**: Ésta es la representación del SDC más antigua y bien establecida, dado que los datos tabulares han sido la forma de publicación tradicional de las agencias nacionales de estadística. El objetivo aquí es publicar información agregada *estática*, como las tablas, de forma que no se pueda deducir información confidencial sobre individuos específicos entre esos a los que se refiere la tabla [9]. Véase [10][11] y [12] para más información.
- **Bases de datos dinámicas**: El escenario aquí es una base de datos a la que el usuario puede enviar consultas estadísticas (sumas, medias, etc.). La información agregada obtenida por los usuarios como resultado de consultas sucesivas no debe permitirles inferir información sobre individuos específicos. Desde los años 80, se sabe que éste es un problema difícil, sujeto al "*tracker attack*" [13]. Las estrategias empleadas actualmente se basan en la perturbación, la restricción o el reemplazo por intervalos de las respuestas a determinadas consultas. Se pueden encontrar ejemplos de esas tres estrategias en [14], [15] y [16] respectivamente.
- **Protección de microdatos**: Esta disciplina se encarga de proteger datos individuales, también llamados microdatos.

En este artículo nos concentramos en la protección de microdatos.

B. Microagregación

La microagregación es una técnica de SDC que consiste en la agregación de microdatos. Recientemente, la tendencia ha sido que los grandes gestores de datos (agencias de estadística u otras entidades por el estilo) publiquen microdatos. Por lo tanto, ésta es la disciplina mas joven dentro de las SDC y está experimentando una evolución continua en los últimos años. Inicialmente fue diseñada para datos numéricos continuos, y recientemente se ha extendido para datos categóricos.

La microagregación se puede entender como un problema de clusterización con restricciones en el tamaño de los clusters. Está relacionada con otros problemas de clusterización (e.g. el problema de reducción de la dimensionalidad). Sin embargo, la diferencia principal del problema de la microagregación es que no considera el número de clusters que se generan o el número de dimensiones que se reducen, sino únicamente el mínimo número de elementos agrupados en el mismo cluster.

Como hemos indicado en la introducción, cuando micro-agregamos datos es necesario tener en cuenta dos objetivos:

- *Preservar la utilidad de los datos*. Para hacerlo, deberíamos introducir el menor ruido posible en los datos, *i.e.* deberíamos agregar elementos similares en lugar de elementos diferentes.
- *Proteger la privacidad de los respondientes*. Los datos tienen que estar suficientemente modificados para hacer que la reidentificación sea difícil, *i.e.* incrementando el número de elementos agregados, incrementamos la privacidad de los datos.

Seguidamente describimos el proceso de microagregación. Antes, es necesario saber que los atributos de un conjunto de datos desprotegido pueden clasificarse en tres categorías:

- *Identificadores:* identifican sin ambigüedad al respondiente (*e.g.* el DNI, el número de pasaporte, el número de la seguridad social, el nombre completo, etc.).
- *Quasi-identificadores:* se trata de un conjunto de atributos que, en combinación, pueden relacionarse con información externa para re-identificar a los respondientes (*e.g.* la edad, el sexo, el empleo, el código postal, etc.).
- *Confidenciales:* contienen información sensible sobre el respondiente (*e.g.* el sueldo, la religión, la afiliación política, si tiene o no una enfermedad, etc.). Éstos son los que tratamos de proteger.

Cuando nos disponemos a microagregar un conjunto de datos, lo primero que debemos hacer es identificar los atributos y suprimir directamente los identificadores. Posteriormente, ejecutamos el algoritmo de microagregación y obtenemos una *k*-partición.

Definición 1 (k-partición): Una *k*-partición es una división del conjunto original de datos en subconjuntos disjuntos de al menos *k* registros.

El último paso del proceso de microagregación es sustituir cada registro del conjunto de datos original por el centroide (i.e. el registro medio) del grupo al que pertenece en la *k*-partición.

C. Método MDAV

Para obtener los resultados experimentales presentados en la Sección IV, hemos utilizado el algoritmo MDAV (Distancia Máxima al Vector Promedio) [22], dado que actualmente es el método más popular para microagregar conjuntos de datos. Este algoritmo proporciona grupos de cardinalidad constante *k* y, en el caso de que el número inicial de registros no sea divisible por *k*, un grupo de cardinalidad entre *k* i 2*k*-1 también se genera (*cf.* Algoritmo 1).

Algoritmo 1 MDAV

X: el conjunto de datos que queremos microagregar.
k: el número mínimo de registros por cluster.
n: el número de registros en el conjunto de datos *X*.
1: ComputarMatrizDistancias(X);
2: registros_restantes = n;
3: i = 0;
4: **mientras** (registros_restantes > 2*k*-1) **hacer**
5: c ← ComputarCentroide(X);
6: r ← RegistroMasLejanoAlCentroide(X, c);
7: s ← RegistroMasLejanoAlRegistro(X, r);
8: g_i ← ConstruirGrupoAlrededorDeRegistro(X, r, *k*);
9: i ++;
10: g_i ← ConstruirGrupoAlrededorDeRegistro(X, s, *k*);
11: i ++;
12: registros_restantes ← registros_restantes – 2*k*;
13: **fin mientras**
14: $g_1...g_s$ ← AssignarRegistrosRestantes(X, $g_1...g_s$);
15: **devuelve** $g_1...g_s$

D. k-Anonimato

El *k*-Anonimato ([17]-[21]) es una aproximación elegante para encarar el conflicto entre la pérdida de información y el riesgo de revelación en SDC.

Definición 2 (k-anonimato): Se dice que un conjunto de datos satisface el *k*-anonimato para $k > 1$ si, para cada combinación de valores de un quasi-identificador, existen al menos *k* registros que comparten esa combinación en el conjunto de datos.

Un método para conseguir *k*-anonimato es la microagregación. En este caso, cada registro original es reemplazado por el centroide del cluster al que pertenece. Así, optener la máxima homogeneidad entre los elementos de un grupo implica minimizar la pérdida de información causada por la microagregación. Hay diversas formas de medir la homogeneidad de los grupos basadas en diferentes definiciones de distancia, siendo la más común para los clusters la Suma de Errores Cuadrados (SEC). Para una *k*-partición dada, el SEC se computa como:

$$SEC = \sum_{i=1}^{s} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)'(x_{ij} - \overline{x}_i) \qquad (1)$$

donde *s* es el numero de clusters en la *k*-partición, n_i es el numero de elementos en el cluster *i*, x_{ij} es el elemento *j* en el cluster *i* y $\overline{x}_i$ es el centroide del cluster *i*.

Los atributos quasi-identificadores son aquellos que utiliza un *snooper* para intentar la reidentificación, *i.e.* relacionar registros en el conjunto de datos protegidos V' con registros identificados en una fuente externa de datos S. Si V' satisface el *k*-anonimato, un *snooper* que intente relacionar registros con S solo puede esperar relacionar un registro identificado de S con un grupo de *k* registros de V' (el *snooper* ve cada registro del grupo como una correspondencia igualmente posible). Por tanto, si, para una *k* dada, se asume que el *k*-anonimato es protección suficiente, uno puede concentrarse en minimizar la pérdida de información con la única restricción de que el *k*-anonimato debe cumplirse.

Cuando se propuso el *k*-anonimato, el procedimiento computacional sugerido para implementarlo fue una combinación de generalizaciones y supresiones. Se ha demostrado recientemente en [22] que la microagregación proporciona un camino computacional más unificado para alcanzar el *k*-anonimato, con las ventajas adicionales sobre las generalizaciones/supresiones que:

i) la microagregación se puede aplicar sobre cualquier tipo de datos (incluyendo datos numéricos);
ii) la microagregación no complica el análisis devolviendo un conjunto de datos *k*-anonimizado con nuevas categorías o datos censurados.
iii) la microagregación no transforma los datos numéricos en datos categóricos.

Hay otras propiedades interesantes que se pueden requerir (e.g., *k*-anonimato *p*-sensible [23], *l*-diversidad [24] o *t*-*closeness* [25]), sin embargo en este artículo nos centramos en la propiedad de *k*-anonimato obtenida mediante microagregación.

III. NUESTRO MÉTODO

En esta sección describimos nuestro método de postprocesado. Su propósito es obtener una *k*-partición con grupos más homogéneos. Dado un grupo g_i de la *k*-partición, el algoritmo intenta reducir su EC (Error Cuadrático, *cf.* Ecuación 2) mientras el SEC de la *k*-partición no aumenta.

$$EC_{g_i} = \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)'(x_{ij} - \overline{x}_i) \qquad (2)$$

El algoritmo recibe como entrada una *k*-partición generada mediante un algoritmo de microagregación (*e.g.* MDAV o V-MDAV), y el conjunto de datos original. El objetivo es redistribuir los registros para disminuir las disimilitudes en los clusters.

En primer lugar, se selecciona el cluster menos heterogéneo (*i.e.* el grupo con mayor EC) y se trata de transformarlo del modo mas apropiado de acuerdo con su cardinalidad. Los procedimientos que proponemos para aplicar a un cluster dado son:

- **Partir el cluster en dos nuevos clusters**: Esta acción sólo se puede realizar cuando el tamaño del cluster es al menos *2k*. De lo contrario, se crearía un cluster con menos de *k* registros. Para partir el cluster se usa un algoritmo inspirado en el método de microagregación MDAV [22] (*cf.* Algoritmo 3). Este procedimiento siempre reduce el SEC de la *k*-partición.
- **Destruir el cluster y reasignar sus registros a otros grupos**: Cuando el número de registros de un cluster es *k*, no se puede quitar ningún registro sin romper completamente el cluster (*e.g.* si se quita un registro, entonces la cardinalidad del grupo original es $k - 1$ y esto no produciría una *k*-partición). Así, todos los registros del cluster deben ser asignados a otros clusters (*cf.* Algoritmo 4).
- **Escoger un registro del cluster y asignarlo a otro grupo**: Si la cardinalidad del cluster esta entre *k* y *2k*, entonces algunos de sus registros se pueden reasignar sin romper el cluster original (*cf.* Algoritmo 5).

Cada vez que se aplica uno de estos cambios el algoritmo determina si ha disminuido la pérdida de información del conjunto de datos producida por la microagregación (*i.e.* el SEC). En caso negativo, el cambio no se aplica y se restaura la configuración de los grupos anterior a él.

Una vez se ha aplicado el método de postprocesado, la homogeneidad de la *k*-partición mejora o, en el peor caso, se mantiene.

El proceso completo se describe en el Algoritmo 2.

Algoritmo 2 Postprocesado de la *k*-partición

G: una *k*-partición $\{g_1, g_2, ..., g_s\}$ donde *s* es el número de clusters.
k: el número mínimo de registros por cluster.
p: el porcentaje de clusters que el algoritmo intentará mejorar antes de acabar.

```
1: i=0
2: mejora=cierto
3: mientras (i < (p*s)/100) y mejora hacer
4:       si (mejora) entonces
5:           ComputarEC(G);
6:           OrdenarClustersPorEC(G);
7:           g ← ClusterConMayorEC(G);
8:           mejora ← falso;
9:           i=1;
10:      fin si
11:      si (card(g) > (2k – 1)) entonces
12:          PartirCluster(g,G,k);
13:          mejora ← cierto;
14:          s ← s+1;     //aumentar el número de clusters
15:      sino si card(g)=k entonces
16:          mejora ← DestruirClusterYReasignar(g,G);
17:          si (mejora) entonces
18:              s ← s – 1;        //disminuir el número de clusters
19:          fin si
20:      sino
21:          mejora ← ReasignarRegistro(g,G);
22:      fin si
23:      si (no mejora) entonces
24:          i ← i + 1;
25:      fin si
26:  fin mientras
27:  devuelve G
```

Algoritmo 3 PartirCluster(*g,G,k*)

g: un grupo de registros
G: una *k*-partición
k: la constante de cardinalidad mínima

```
1: c ← Centroide(g)
2: a ← RegistroMasLejanoAlCentroide(c,g)
3: b ← RegistroMasLejanoAlRegistro(a,g)
4: g_a ← Construir un cluster con a y sus registros más
        próximos en g
5: g_b ← Construir un cluster con b y sus registros más
        próximos en g
6: mientras (Card(g_a) < Card(g_b) o Card(g_b) < k) hacer
7:     Asignar al grupo más pequeño el registro del grupo más grande que esté
       más cerca.
8: fin mientras
```

Algoritmo 4 DestruirClusterYReasignarRegistros(g,G)

g: un grupo de registros
G: una *k*-partición

```
1: G' ← Construir una k-partición alternativa en la que g se ha
        destruido y sus registros se han asignado individualmente a sus
        clusters más cercanos.
2: si (SEC(G) > SEC(G')) entonces
3:     G ← G'
4:     devuelve cierto
5: sino
6:     devuelve falso
7: fin si
```

Algoritmo 5 ReasignarRegistro(g,G)

Requiere:
g: un grupo de registros
G: una *k*-partición

```
1: c ← Centroide(g).
2: g_a ← registro más lejano a c en g.
3: nc ← cluster más próximo a g en la k-partición G.
4: Formar un cluster g' con todos los registros en g menos g_a.
5: Formar un cluster nc' con todos los registros en nc menos
   g_a.
6: si (SE(g) + SE(nc)) > (SE(g') + SE(nc')) entonces
7:     Reemplazar g por g'.
8:     Reemplazar nc por nc'.
9: fin si
```

IV. RESULTADOS

En esta sección presentamos algunos resultados para mostrar cómo nuestro algoritmo mejora con respecto del método de Máxima Distancia al Vector Promedio (MDAV). Esta heurística fue introducida por primera vez en el paquete informático µ-Argus para el control de revelación estadística y se describe en [22],[26].

Con la intención de testear nuestro algoritmo de postprocesado, hemos aplicado el método MDAV sobre datos sintéticos y datos reales.

Hemos generado dos tipos de conjuntos de datos sintéticos:

- *Conjuntos de datos agrupados (clustered)*: se trata de conjuntos de datos formados por 100 registros de 2 atributos cada uno. Los registros forman clusters de forma natural con una cardinalidad de entre 3 y 7 registros por grupo. Los valores varían entre -10.000 y 10.000 (*cf.* Figura 1 para un ejemplo de este tipo de conjunto de datos).
- *Conjuntos de datos dispersos (scattered)*: se trata de conjuntos de datos formados por 100 registros de 2 atributos cada uno. Los atributos se obtuvieron a partir de una distribución Gausiana, dando valores desde 10.000 hasta -10.000. En este caso, aparentemente no existen clusters naturales (*cf.* Figura 2 para un ejemplo de este tipo de conjunto de datos).

Para cada tipo se han creado diez conjuntos de datos diferentes para poder obtener resultados promedios.

Por otro lado, hemos utilizado también tres conjuntos conocidos de datos reales, de referencia a la hora de testear métodos de microagregación multivariante [28]:

- *CENSUS:* se trata de un conjunto de datos formado por 1080 registros de 13 dimensiones que contiene información cedida por la Oficina del Censo Norteamericana.
- *EIA:* se trata de un conjunto de datos formado por 4092 registros de 11 dimensiones con información cedida por la Agencia Internacional de la Energía.
- *Tarragona:* se trata de un conjunto de datos formado por 834 registros y 13 dimensiones cedido por la Cámara de comercio de Tarragona.

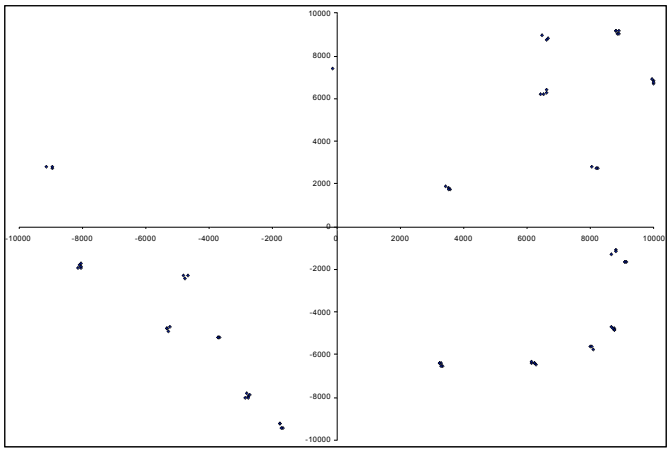


Fig. 1. Ejemplo de un conjunto de datos *clustered* usado en los experimentos.

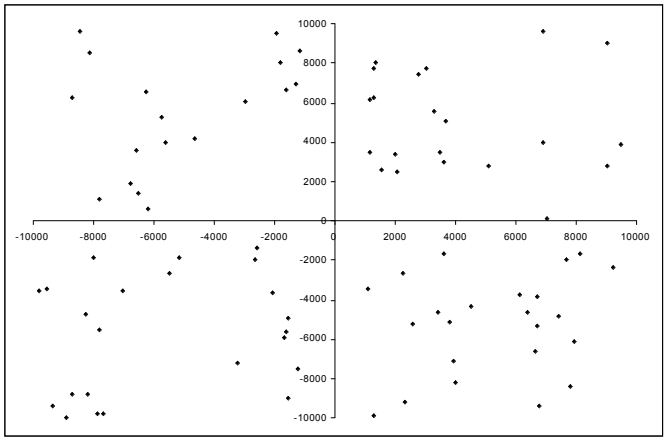


Fig. 2. Ejemplo de un conjunto de datos *scattered* usado en los experimentos.

Después de haber obtenido una *k*-partición para cada conjunto de datos por medio del método MDAV, hemos aplicado el Algoritmo 2 para observar las mejoras en las *k*-particiones en términos de SEC. El parámetro *p* (*i.e.* el porcentaje de clusters que el método intenta mejorar) se ha establecido para efectuar el postprocesado sobre todos los clusters (*i.e.* 100%). En todos los experimentos que se comentan a continuación, el valor de *k* es 3.

Para mostrar el rendimiento del algoritmo de postprocesado medimos el porcentaje de mejora del SEC. Para hacerlo usamos la siguiente expresión:

$$IMP = \frac{SEC_{MDAV} - SEC_{post}}{SEC_{MDAV}} \times 100 \qquad (3)$$

donde SEC_{MDAV} es el SEC de la *k*-partición obtenido usando el método MDAV y SEC_{post} es el SEC de la *k*-partición obtenido después de postprocesar con nuestro método la *k*-partición obtenida con el MDAV.

Observando las Figuras 3 y 4 se ve claramente que la aplicación de nuestro algoritmo de postprocesado provoca un decremento del SEC (*i.e.* las *k*-particiones obtenidas producen conjuntos de datos *k*-anónimos con menos pérdida de información).

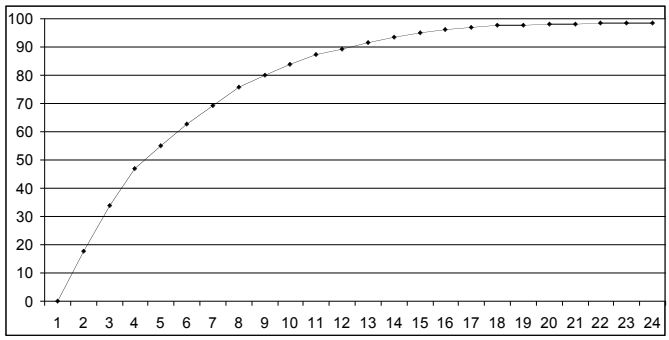


Fig. 3. Porcentaje de mejora de los conjuntos de datos *clustered* (media sobre diez conjuntos).

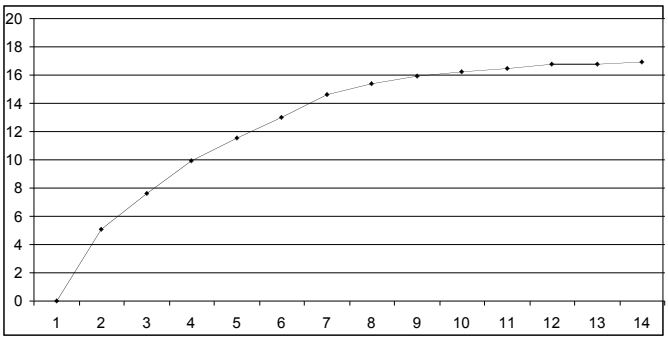


Fig. 4. Porcentaje de mejora de los conjuntos de datos *scattered* (media sobre diez conjuntos).

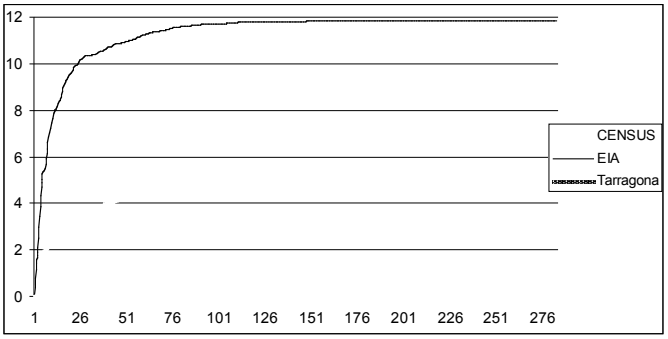


Fig. 5. Porcentaje de mejora de los conjuntos de datos CENSUS, EIA y Tarragona.

Aunque hemos comprobado que el método propuesto es útil en todos los tipos de conjuntos de datos testeados, sintéticos y reales, su rendimiento es claramente mejor cuando se aplica sobre conjuntos de datos *clustered*. Para éstos, nuestra propuesta mejora significativamente la *k*-partición obtenida por el MDAV, dado que el SEC final es un 99% menor que el original. La Figura 3 muestra la mejora (IMP), definida en la Expresión 3, a cada paso de ejecución. De acuerdo con [26], el método MDAV es una heurística de tamaño fijo (*i.e.* lo que resulta en una *k*-partición con clusters formados por *k* registros), y por lo tanto proporciona un resultado pobre cuando se aplica a conjuntos de datos *clustered*. Nuestro método de postprocesado mejora claramente los resultados del MDAV en términos de SEC porque es capaz de generar clusters de cardinalidad variable entre k y $2k - 1$.

Por otro lado, aunque MDAV ya obtiene un rendimiento bueno sobre los conjuntos de datos *scattered*, nuestro método de postprocesado es capaz de disminuir en un 16% el SEC de estos conjuntos de datos, como muestra la Figura 4.

En la Figura 5 podemos observar que en los conjuntos de datos reales obtenemos una mejora claramente distinta en según el conjunto de datos: mientras que el conjunto EIA mejora en un 12%, el conjunto Tarragona no llega al 0.1%.

V. CONCLUSIÓN

Hemos presentado un método de postprocesado para disminuir los errores introducidos por los métodos de SDC. Hemos mostrado que usando nuestra propuesta la Suma de Errores Cuadrados se puede reducir al mismo tiempo que se mantiene la privacidad de los respondientes.

El método que hemos expuesto es este artículo es útil para reducir los errores introducidos en datos sintéticos y reales por los métodos de SDC. Sin embargo, en un futuro se harán diversas mejoras y experimentos:

- Valorar la utilidad de nuestro método con otras técnicas de SDC. En este artículo hemos aplicado nuestro método de postprocesado a las *k*-particiones obtenidas por el método MDAV. Testearemos nuestro método en las *k*-particiones obtenidas por otras técnicas de SDC como el V-MDAV [27] o la µ-aproximación [22].
- Extender nuestro método para tratar con las nuevas propiedades como el *k*-anonimato *p*-sensible [23], la *l*-diversidad [24] o *t*-cercanía [25].
- Incrementar el número de primitivas que puede utilizar el algoritmo: hemos introducido tres primitivas principales (*i.e.* partir el cluster en dos, destruir el cluster y reasignar, reasignar un registro del cluster). Sin embargo, se pueden añadir diversas primitivas nuevas al método (*e.g.* intercambiar registros, reasignar varios registros a la vez).
- Analizar el coste computacional del método propuesto.

AGRADECIMIENTOS

Los autores forman parte de la Cátedra UNESCO de Privacidad de Datos, pero sus opiniones no necesariamente reflejan la posición de la UNESCO ni comprometen a dicha organización. Este trabajo ha sido parcialmente subvencionado por el Ministerio Español de Educación y Ciencia mediante los proyectos TSI2007-65406-C03-01 "E-AEGIS" y CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", y por la Generalitat de Catalunya mediante la subvención 2005 SGR00446.

REFERENCIAS

[1] HIPAA, .Health insurance portability y accountability act,. 2004, http://www.hhs.gov/ocr/hipaa/.

[2] C. Boyens, R. Krishnan, y R. Padman, "On privacypreserving access to distributed heterogeneous healthcare information", en *Proceedings of the 37th Hawaii International Conference on System Sciences HICSS-37*, I. C. Society, Ed., Big Isly, HI., 2004.

[3] A. Solanas y A. Martínez-Ballesté, "Privacy protection in location-based services through a public-key privacy homomorphism", en *4th European PKI Workshop. EuroPKI'07*, J. Lopez, P. Samarati, y J. Ferrer, Eds., vol. 4582. Springer, June 2007, pp. 362 . 368.

[4] A. Solanas, J. Domingo-Ferrer, A. Martínez-Ballesté, y V. Daza, "A Distributed Architecture for Scalable RFID Identification", *Computer Networks*, vol. 51, 2007.

[5] European Parliament, "DIRECTIVE 2002/58/EC of the european parliament and council of 12 july 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications)", 2002, http://europa.eu.ent/eur-lex/pri/en/oj/dat/2002/l 201/l 20120020731en00370047.pdf.

[6] Canadian Privacy, "Canadian privacy regulations", 2005, http://www.media-awareness.ca/english/issues/privacy/canadian legislation privacy.cfm.

[7] U.S. Privacy, "U.S. privacy regulations", 2005, http://www.media-awareness.ca/english/issues/privacy/us legislation privacy.cfm.

[8] A. Oganian y J. Domingo-Ferrer, "On the complexity of optimal microaggregation for statistical disclosure control", *Statistical Journal*

of the United Nations Economic Comission for Europe, vol. 18, no. 4, pp. 345. 354, 2001.

[9] J. Castro, "Minimum-distance controlled perturbation methods for large-scale tabular data protection", *European Journal of Operational Research*, vol. 171, no. 1, pp. 39.52, May 2006.

[10] L. Willenborg y T. DeWaal, *Elements of Statistical Disclosure Control*. Springer-Verlag, New York, 2001.

[11] S. Giessing, "Survey on methods for tabular data protection in ARGUS", en *Privacy in Statistical Databases*, ser. Lecture Notes in Computer Science, J. Domingo-Ferrer y V. Torra, Eds., vol. 3050. Berlin Heidelberg: Springer, 2004, pp. 1.13.

[12] J. Domingo-Ferrer y J. M. Mateo-Sanz, "On resampling for statistical confidentiality in contingency tables", *Computers & Mathematics with Applications*, no. 38, pp. 13.32, 1999.

[13] J. Schlörer, "Disclosure from statistical databases: quantitative aspects of trackers", *ACM Transactions on Database Systems*, vol. 5, pp. 467.492, 1980.

[14] F. Y. Chin y G. Ozsoyoglu, "Auditing and inference control in statistical databases", *IEEE Transactions on Software Engineering*, no. 8, pp. 574.582, 1982.

[15] G. Duncan y S. Mukherjee, "Optimal disclosure limitation strategy in statistical databases: deterring tracker attacks through additive noise", *Journal of the American Statistical Association*, no. 95, pp. 720.729, 2000.

[16] R. Garfinkel, R. Gopal, y D. Rice, "New approaches to disclosure limitation while answering queries to a database: protecting numerical confidential data against insider threat based on data and algorithms", 2004, manuscript.

[17] J. Domingo-Ferrer, A. Solanas y A. Martínez-Ballesté, "Privacy in statistical databases: k-anonymity through microaggregation", en *IEEE Granular Computing'06*. Atlanta. USA: IEEE, May 2006.

[18] L. Sweeney, "k-anonymity: A model for protecting privacy", *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, vol. 10, no. 5, pp. 557. 570, 2002.

[19] P. Samarati y L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization y suppression", SRI International, Tech. Rep., 1998.

[20] P. Samarati, "Protecting respondents' identities in microdata release", *IEEE Transactions on Knowledge y Data Engineering*, vol. 13, no. 6, pp. 1010.1027, 2001.

[21] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression", *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, vol. 10, no. 5, pp. 571.588, 2002.

[22] J. Domingo-Ferrer, F. Sebe, y A. Solanas, "A poliynomial-time approximation to optimal multivariate microaggreagation", *Computers y Mathematics with Applications (to appear)*, 2007.

[23] T. M. Truta y B. Vinay, "Privacy protection: p-sensitive k-anonymity property", en *2nd International Workshop on Privacy Data Management PDM*. Berlin Heidelberg, 2006. IEEE Computer Society., 2006, p. 94.

[24] D. K. A. Machanavajjhala, J. Gehrke y M. Venkitasubramanian, "l-diversity: privacy beyond k-anonymity", en *Proceedings of the IEEE ICDE 2006*, 2006.

[25] T. L. N. Li y S. Venkatasubramanian, "T-closeness: privacy beyond k-anonymity and l-diversity," en *Proceedings of the IEEE ICDE 2007*, 2007.

[26] J. Domingo-Ferrer, A. Martínez-Ballesté, J. M. Mateo-Sanz, y F. Sebé, "Efficient multivariate data-oriented microaggregation". *The VLDB Journal*, vol. 15, no. 4, pp. 355.369, 2006.

[27] A. Solanas y A. Martínez-Ballesté, "V-MDAV: Variable group size multivariate microaggregation", en *COMPSTAT'2006*, Rome, 2006, pp. 917.925.

[28] R. Brand, J. Domingo-Ferrer, J. M. Mateo Sanz. "Reference data sets to test and compare SDC methods for protection of numerical microdata", 2002. European Project IST-2000-25069 CASC.



Glòria Pujol obtuvo la Ingeniería Técnica en Informática de Sistemas (2007) en la Universidad Rovira y Virgili (Av. Països Catalans 26, 43007 Tarragona).

En la actualidad es becaria de proyecto en el grupo de investigación CRISES de la Universidad Rovira y Virgili (Av. Països Catalans 26, 43007 Tarragona). Sus intereses de investigación están relacionados con la privacidad y seguridad de datos.



Antoni Martínez-Ballesté obtuvo la Ingeniería Técnica en Informática de Sistemas (1999) y la Ingeniería Superior en Informática (2002) en la Universidad Rovira y Virgili (Av. Països Catalans 26, 43007 Tarragona). En 2004 recibió con honores el doctorado en Ingeniería Telemática por la Universidad Politécnica de Catalunya (C/ Jordi Girona, 31. 08034 Barcelona).

En 2004, fue investigador visitante en LAAS-CNRS, Toulouse, Francia. En la actualidad imparte docencia en la Universidad Rovira y Virgili (Av. Països Catalans 26, 43007 Tarragona). Ha publicado más de 30 artículos en revistas y conferencias. Sus intereses de investigación están relacionados con la seguridad en las redes de computadoras y las aplicaciones seguras.



Josep M. Mateo-Sanz obtuvo la licenciatura en Ciencias Matemáticas (1992) y el Doctorado en Ciencias Matemáticas (1998) en la Universidad de Barcelona (Gran Via de les Corts Catalanes 585, 08007 Barcelona). Es también diplomado en Profesorado de Educación General Básica (1985).

En la actualidad es Titular de Universidad del área de Estadística y Investigación Operativa en la Universidad Rovira y Virgili (Av. Països Catalans 26, 43007 Tarragona). Es autor de diez artículos internacionales, uno de los cuales fue ISI Highly Cited Paper a principios del 2005. Sus intereses de investigación son el secreto estadístico, la privacidad y seguridad de los datos y la aplicación de la estadística en diversas áreas del conocimiento.



Agusti Solanas obtuvo la Ingeniería Técnica en Informática de Sistemas (2002) por la Universidad Rovira i Virgili (URV) de Tarragona. En 2004 obtuvo el título de Ingeniero en Informática por la URV graduándose como primero de la promoción. En 2007 recibió con honores el doctorado en Ingeniería Telemática por la Universidad Politécnica de Cataluña (UPC). Fue seleccionado para tomar parte en el Premio Jóvenes Investigadores 2005 organizado por el Ministerio Español de Educación, Cultura y Deporte. En la actualidad es investigador del grupo CRISES de la URV y miembro de la Cátedra UNESCO en Privacidad de Datos. Ha participado en diversos proyectos de investigación nacionales e internacionales, ha publicado más de 30 artículos y dado varias conferencias. También ha sido miembro de diversos comités científicos y revisor en conferencias y revistas. Sus intereses de investigación están relacionados con la privacidad y la seguridad, la tecnología RFID, los servicios basados en localización (LBS) y la inteligencia artificial. Es miembro de diversas asociaciones y sociedades científicas, entre ellas: la IACR (International Association of Criptologic Research), el IEEE (Institute of Electrical and Electronics Engineers), la ACM (Association for Computing Machinery), la Sigma Xi (The Scientific Research Society).