Publicación Electrónica **ISSN 2718- 6636**

UNIVERSIDAD NACIONAL DE ROSARIO
FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA
SECRETARIA DE CIENCIA Y TECNOLOGIA E INSTITUTOS DE INVESTIGACIONES

Resumen Ampliado

Jornadas Anuales
"Investigaciones en la Facultad"
Ciencias Económicas y Estadística

Bussi, Javier[*]; Wojdyla, Daniel M.[**]
[*]Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística, Facultad de Ciencias Económicas y Estadística, Universidad Nacional de Rosario
[**]Duke Clinical Research Institute, Duke University

UTILIDAD DE LOS DATOS Y SU RELACIÓN CON DISTINTOS MÉTODOS DE ANONIMIZACIÓN UTILIZANDO EL PROPENSITY SCORE[1]

Resumen
En los últimos años ha crecido la idea de que los datos deben estar disponibles, de manera accesible y rápida, al conjunto de la sociedad. En la Argentina, existe la obligación de respetar el secreto estadístico de las unidades de información y por lo tanto es necesario desarrollar alternativas que permitan satisfacer ambas metas: publicar los datos y mantener el secreto estadístico. Es por este motivo que se estudian distintos métodos de anomización que se engloban dentro de lo que se denomina el Control Estadístico de Divulgación. El objetivo de los métodos de anonimización es minimizar el riesgo de divulgación de la identidad de las unidades de información manteniendo la utilidad de los datos, evaluada a través de distintas medidas basadas en la distancia entre los datos originales y perturbados. En este trabajo se presentan dos medidas que utilizan el Propensity Score y se muestran sus desempeños en un conjunto de datos provenientes de Estadísticas Oficiales. Se realizó un estudio de simulación considerando la combinación de dos métodos de anonimización (microagregación y ruido corelacionado), las dos medidas de utilidad propuestas y 3 niveles de riesgo de divulgación. Este estudio muestra que ambas medidas son sensibles a distintos grados de anonimización: sus valores crecen a medida que la utilidad de los datos disminuye. Las mismas proveen una medida global de utilidad apropiada para evaluar la pérdida de información.

Palabras clave: Control Estadístico de Divulgación, Estadísticas Oficiales, Propensity Score

Abstract
In recent years, the idea that data must be available to the general public easily and quickly has gained momentum. In Argentina, there is an obligation to respect the confidentiality of the information units and therefore it is essential to develop methods that allow both goals to be met: publishing the data and maintaining statistical secrecy. It is for this reason that different anomization methods are studied that are included within what is called Statistical Disclosure Control. The objective of the anonymization methods is to minimize the risk of disclosure of the identity of the units of information while maintaining the utility of the data that is evaluated through different measures that consider the distance between the original and the modified data. In this work, two of the measures based on the Propensity Score are presented and their performance is shown in a dataset from Official Statistics. A simulation study was carried out considering the combination of two anonymization methods (microaggregation and correlated noise), the two proposed utility measures and 3 levels of disclosure risk. This study shows that both measures presented are sensitive to different degrees of anonymization, since their values

[1] Trabajo elaborado en el marco del Proyecto (80120210200167UR), titulado: “Métodos muestrales y series de tiempo en las estadísticas oficiales”, dirigido por Mg. Gonzalo P.D. Marí.

increase as the utility of the data decreases. They provide an appropriate global measure of utility to evaluate information loss.

Keywords: Statistical Disclosure Control, Official Statistics, Propensity Score

Introducción

Las Estadísticas Oficiales proveen información sobre distintos aspectos de la realidad de un país, referidas a personas, hogares o empresas, solo para mencionar algunos ejemplos. Permiten caracterizar el estado de situación y planificar en base a la toma de decisiones. En los últimos años ha crecido la idea de que los datos deben estar disponibles, de manera accesible y rápida, al conjunto de la sociedad para que puedan ser analizados, y que los mismos deben ser gratuitos, reutilizables y sin restricciones referidas a derechos de autor. En la Argentina, existe la Ley N° 17.622 que establece la obligación de respetar el secreto estadístico de las unidades de información. Esto supone un desafío a la hora de publicar datos oficiales que podrían ser solucionado a través de diferentes acciones:

a) Publicar bases para usuarios sin información acerca de la identificación de las unidades ni de las variables de diseño muestral pero esto no permitiría realizar análisis apropiados de los datos
b) Modificar los valores de ciertas variables consideradas claves antes de publicarlos, pero se correría el riesgo de distorsionar las estimaciones en los análisis realizados
c) No publicar la información, es una medida segura pero que se contrapone con la política de datos abiertos y es inaceptable para el conjunto de la sociedad.

Por lo tanto, es necesario desarrollar métodos que permitan satisfacer ambas metas: publicar los datos y mantener el secreto estadístico. Es por este motivo que se estudian distintos métodos de anomización que se engloban dentro de lo que se denomina el Control Estadístico de Divulgación. El objetivo de los métodos de anonimización es minimizar el riesgo de divulgación de la identidad de las unidades de información manteniendo la utilidad de los datos. Existen distintas medidas para evaluar la utilidad de los datos que consideran la distancia entre los datos originales y perturbados, en este trabajo se presenta dos medidas basadas en el Propensity Score mostrando su desempeño en un conjunto de datos provenientes de Estadísticas Oficiales.

Objetivo

Proveer una medida global de utilidad apropiada para evaluar la pérdida de información en el Control Estadístico de Divulgación utilizando el Propensity Score.

Metodología y materiales

El objetivo de la anonimización en las Estadísticas Oficiales es minimizar el riesgo de divulgación preservando la utilidad de los datos. Existen distintas medidas para evaluar esta cuestión, en este trabajo se presenta dos medidas de utilidad de los datos basadas en el Propensity Score (Woo et al, 2009). Este método es frecuentemente utilizado en estudios médicos observacionales donde el Propensity Score es la probabilidad de un individuo de recibir un tratamiento dado ciertas covariables. En esta aplicación, se compara la distribución de estos scores entre los datos originales y los anonimizados. Si ambos grupos cuentan con

una distribución similar en cuanto al Propensity Score, las variables anonimizadas contarán con una distribución similar a las variables originales. El procedimiento contempla las siguientes instancias, en primer lugar, se combinan los datos originales y los datos modificados por algún método de anonimización. Se crea una variable indicadora que asigne el valor 1 a los datos anonimizados y 0 a los datos originales. A partir de esta estructura de datos se estima un modelo de regresión logística donde la variable respuesta está constituida por la variable indicadora creada y las variables explicativas son las variables anonimizadas, que están conformadas en parte por los valores originales y en parte por los valores perturbados. Las proporciones estimadas por el modelo se comparan con la proporción de datos anonimizados (C) en las variables utilizadas como explicativas en el modelo, que en general es igual a ½.
En este trabajo se consideraron dos técnicas de perturbación para variables continuas: adición de ruido correlacionado y microagregación.
En el caso de ruido correlacionado, siendo $\boldsymbol{X}$ la matriz de datos originales para las p variables con media $\boldsymbol{\mu}$ y matriz de covariancias $\boldsymbol{\Sigma}$, se definen las variables perturbadas $\boldsymbol{Z}$ de la siguiente manera:

$$\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{\epsilon}$$

donde $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\epsilon})$ siendo $\boldsymbol{\Sigma}_{\epsilon} = c\boldsymbol{\Sigma}$ lo que implica que $\boldsymbol{\Sigma}_{Z} = \boldsymbol{\Sigma} + \mathrm{c}\boldsymbol{\Sigma} = (1 + \mathrm{c})\boldsymbol{\Sigma}$
En el caso de microagregación, el método particiona los datos en grupos de k vecinos más cercanos, luego asigna algún valor a cada observación en el grupo, por ejemplo, el promedio, y para mantener la estructura multivariada de los datos este agrupamiento se hace de acuerdo a una medida de similaridad, de acuerdo a distintos criterios. En este trabajo se utiliza la Máxima Distancia al Valor Medio (MDAV), para k=3 y k=5 vecinos más cercanos reemplazando los valores en cada grupo por la media aritmética de la variable correspondiente.
Para reducir el riesgo de divulgación se combinaron ambos métodos adicionándose ruido correlacionado a la microagregación para alcanzar un determinado nivel de riesgo.
El riesgo considerado en este trabajo fue el de divulgación por intervalo: se define como el porcentaje de observaciones originales que caen en el intervalo calculado sobre las observaciones perturbadas teniendo en cuenta una medida de dispersión, por ejemplo, el desvío estándar.
Se consideraron las siguientes medidas de utilidad basadas en el Propensity Score:

a) UP: basada en las diferencias al cuadrado entre los valores predichos y la proporción de datos anonimizados (Woo et al,2009)

$$UP = \frac{1}{n_m} \sum_{i=1}^{n_m} (p_i - c)^2$$

donde $n_m$es el número de observaciones en los datos combinados, $p_i$son las predicciones hechas con el modelo de regresión logística y c es la proporción de datos anonimizados en el conjunto de datos combinados, en general $c = 1/2$. Si UP toma un valor cercano a 0, la utilidad de los datos será alta, si toma un valor cercano a ¼ , los datos tendrán baja utilidad. Un valor bajo implica que los datos no son distinguibles, y el valor máximo que son perfectamente distinguibles.

b) UPA: basada en las diferencias absolutas entre los valores predichos y la proporción de datos anonimizados.

$$UPA = \frac{1}{n_m} \sum_{i=1}^{n_m} |p_i - c|$$

donde n_m, p_i y c se definen como en el punto anterior. Si UPA toma un valor cercano a 0, la utilidad de los datos será alta, si toma un valor cercano a 1/2, los datos tendrán

baja utilidad. Un valor bajo implica que los datos no son distinguibles, y el valor máximo que son perfectamente distinguibles.

Para la aplicación se consideró Encuesta Nacional de Gasto de los Hogares 2017/18 (ENGHO) para la población urbana (2.000 y más habitantes) de la provincia de Buenos Aires. La base usuario utilizada se encuentra disponible en https://www.indec.gob.ar/indec/web/Institucional-Indec-BasesDeDatos-4 . Se consideraron las variables continuas Ingreso Total del Hogar (ingtoth) y Gasto Total del Hogar (gastot), la variable discreta Cantidad de miembros del hogar (cantmiem) y la variable categórica Propiedad de automóvil (propauto). Se estimó un modelo de regresión logística considerando como variables explicativas estas cuatro variables y la interacción de las variables continuas. La variable respuesta fue la variable indicadora de la anonimización y se computaron los propensity scores. Se realizó un estudio de simulación considerando los métodos de anonimización presentados, las dos medidas de utilidad propuestas y el control del riesgo de divulgación considerando 3 niveles. Se plantearon 6 escenarios considerando la adición de ruido correlacionado a microagregación con k=3 y k=5 vecinos más cercanos, controlando el riesgo de divulgación: 75 %, 50 % y 25 %. Cada escenario contó con 100 repeticiones. Se promediaron los valores obtenidos en cada repetición de las medidas UP y UPA. Estos promedios fueron comparados entre los distintos riesgos de divulgación. Para los cálculos se utilizó el programa R y en particular para la perturbación de las variables se utilizó el paquete sdcMicro del mismo programa.

Resultados

En la Figura 1 se presentan los resultados para microagregación con k=3 con adición de ruido correlacionado hasta alcanzar el nivel de riesgo de divulgación estipulado. Tanto UP como UPA muestran valores más altos para menores riesgos de divulgación indicando una menor utilidad de los datos anonimizados. Se puede observar que UP exhibe una mayor variación relativa a través de los distintos niveles de riesgo de divulgación siendo $UP25\%$ un 21% más alto que UP 50% versus UPA 25% un 9% más alto que UPA 50%. Este resultado no es sorprendente debido a las distintas escalas de ambas medidas. Aún así, UP pareciera ser más útil por una simple cuestión descriptiva, del nivel de utilidad de los datos, al mostrar cambios relativos más importantes para distintos niveles de riesgo.

Similarmente, en la Figura 2 se presentan los resultados para microagregación con k=5 con ruido correlacionado adicionado, donde se observa el mismo comportamiento que el caso de microagregación con k=3 aunque la variación relativa para cada medida propuesta entre los distintos riesgos de divulgación es menor.

Figura 1: Microagregación (k=3) más ruido correlacionado.

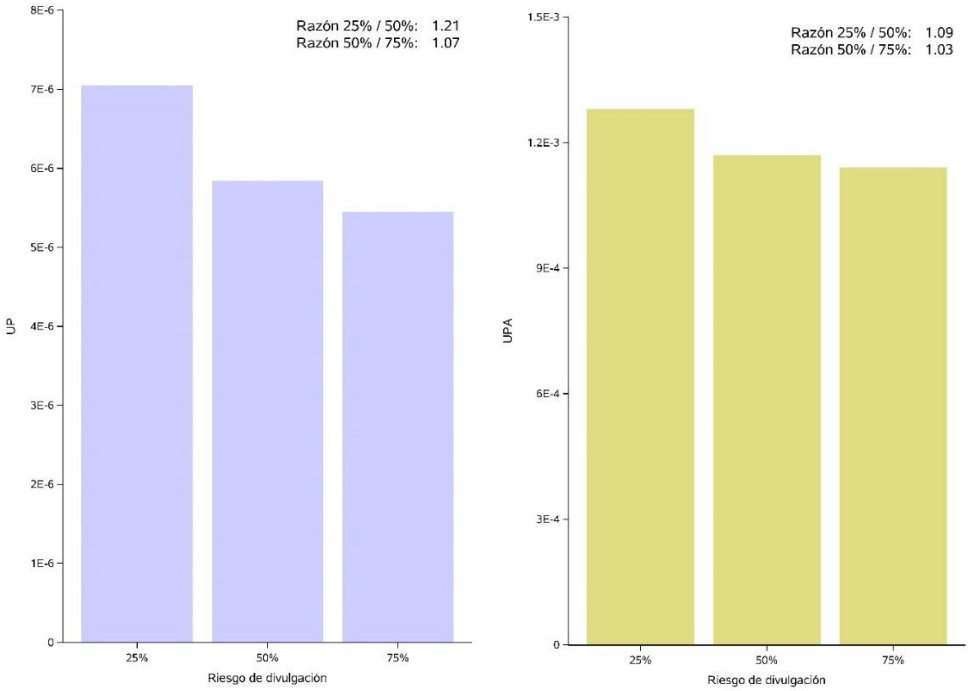
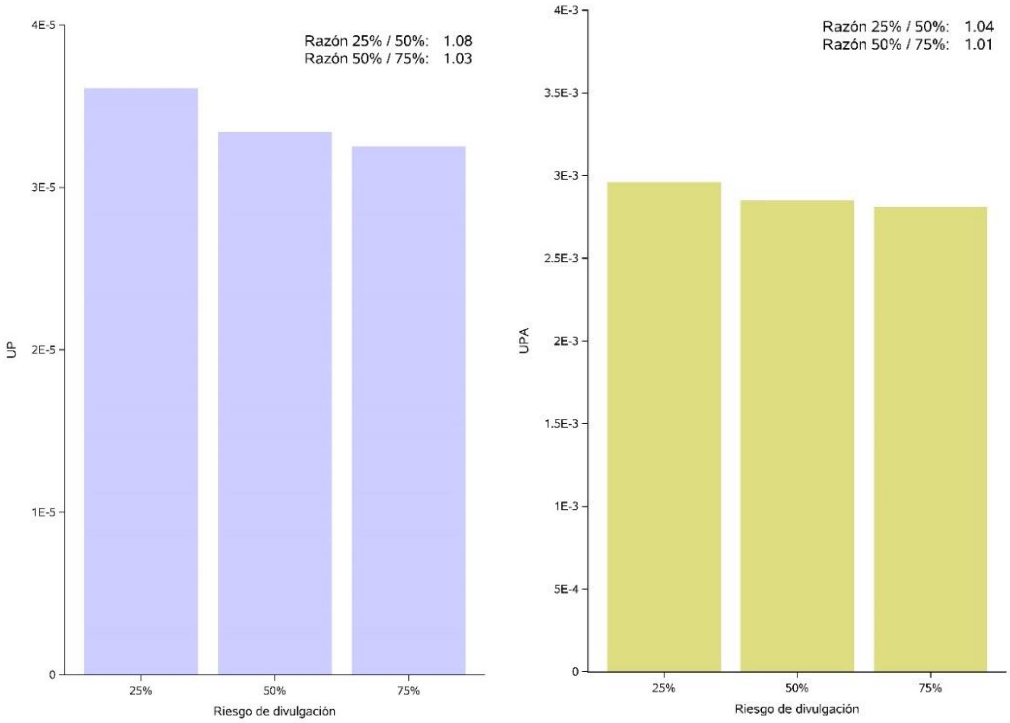


Figura 2: Microagregación (k=5) más ruido correlacionado.



Al comparar los distintos niveles de microagregación se observa que tanto en UP (Figura 3) como en UPA (Figura 4), a mayor número de vecinos microagregados mayor resulta la medida de utilidad, mostrando menor utilidad de los datos cuando se promedian mayor cantidad de observaciones.

Figura 3: UP en microagregación k=3 vs. K=5 con ruido aditivo correlacionado

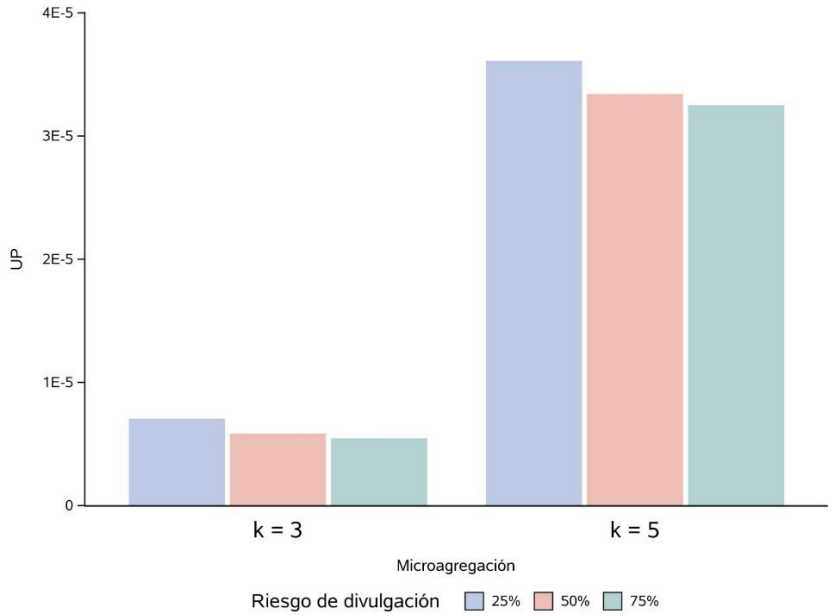
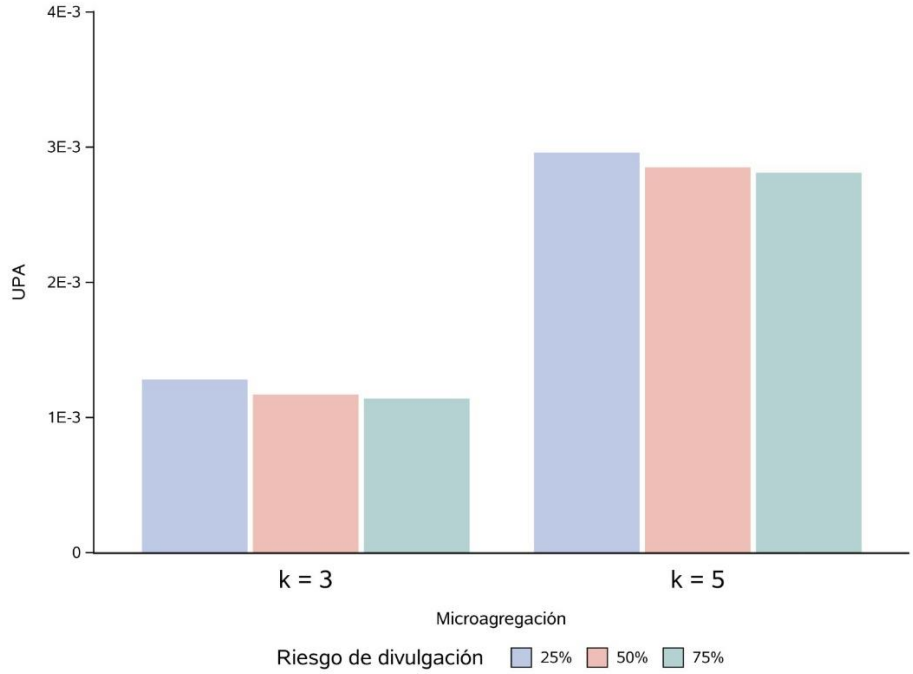


Figura 4: UPA en microagregación k=3 vs. K=5 con ruido aditivo correlacionado



Conclusiones

Se puede observar en este estudio que ambas medidas UP y UPA son sensibles a los distintos grados de anonimización, ya que sus valores crecen a medida que se pierde utilidad en los datos. Con el fin de lograr menores riesgos de divulgación es necesario mayor perturbación de las observaciones, y esto produce que estas medidas aumenten su valor. Las mismas proveen una medida global de utilidad apropiada para evaluar la pérdida de información.

Discusión

Al promediar más observaciones en microagregación, se generan valores más altos de estas medidas de utilidad, en consecuencia, se debe utilizar con cautela este método ya que genera una pseudo categorización de las variables continuas. La microagregación parece tener un efecto más profundo que la adición de ruido en la utilidad de los datos. Como próximo estudio resultaría interesante evaluar cómo se comportan estas medidas ante la inclusión de variables categóricas modificadas y el impacto de estimar distintos modelos de regresión logística para el cálculo del propensity score.

REFERENCIAS BIBLIOGRÁFICAS

- Statistical Disclosure Control for Microdata. Methods and Applications in R. Templ, M. Springer.(2017)

- Global Measures of Data Utility for Microdata Masked for Disclosure Limitation. The Journal of Privacy and Confidentiality. Woo et. al. (2009)

- Assessing, visualizing and improving the utility of synthetic data. Expert Meeting on Statistical Data Confidentiality. Raab et.al. (2021)

- Synthetic Data: An Exploration of Data Utility and Disclosure Risk. PhD Thesis. Taub, J. (2020)