Publicación Electrónica ISSN 1668-5008

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

ACTAS

Jornadas Anuales

"Investigaciones en la Facultad" de Ciencias Económicas y Estadística

Bussi, Javier
Marí, Gonzalo
Méndez, Fernanda

Instituto de investigación, Escuela de Estadística

INDICADORES DE CALIDAD DE LA INFORMACIÓN EN LA ANONIMIZACIÓN DE BASES DE DATOS DE ESTADÍSTICAS OFICIALES[1]

Resumen:

Los operativos estadísticos tienen por objetivo recolectar datos de unidades de observación que pueden estar constituidas por personas, hogares, empresas u otros objetos. Estos datos proveen información que, para el caso de las estadísticas oficiales, proporcionan elementos importantes para la evaluación de políticas implementadas y para la toma de decisiones a futuro. Si bien este es el objetivo primario, no puede desconocerse que existe una marcada tendencia en aumento a la difusión de la información a la sociedad para que la misma utilice los datos para analizarlos y poder obtener sus propias conclusiones. De esta forma, en los últimos tiempos, el término datos abiertos tuvo una expansión significativa, al punto tal que existen portales oficiales que contienen datos a los cuales el público en general puede acceder. Por otra parte, la Ley N° 17.622 menciona la obligación de respetar el secreto estadístico en la divulgación de los operativos que se llevan a cabo. De esta forma, existe una contraposición entre ambas posturas, datos abiertos versus secreto estadístico, que es importante a tener en cuenta para buscar soluciones que permitan brindar información sin violar el secreto estadístico. Es por este motivo, que se estudian distintas alternativas que se engloban dentro del término anonimización. Se propone utilizar métodos de anonimización perturbadores, los cuales modifican los datos observados de las unidades para evitar que un intruso pueda detectar, con información secundaria, la identidad de la misma, lo que provocaría que el organismo que difunde la información violase la ley mencionada. Se evalúan las técnicas de perturbación que adicionan un ruido aleatorio a los datos y la de microagregación, que asignan un valor representativo a todas las unidades que forman parte de un grupo de unidades determinado por la cercanía de las mismas. Se comparan los métodos propuestos a través de medidas que dan cuenta de la distorsión que producen en las estimaciones de ciertos parámetros la aplicación de los métodos de perturbación a los datos originales. En el presente trabajo, se evalúan ciertos escenarios de anonimización sobre la base usuario de la ENGHO publicada por el INDEC, utilizando las bases de pesos replicados para la estimación de ciertos parámetros que cuantifican el error muestral. En general, los métodos que adicionan un ruido aleatorio no correlacionado presentan mejores resultados que sus competidores, y más consistentes que los que produce el método de microagregación.

[1] Trabajo elaborado en el marco del Proyecto (ECO 1ECO199), titulado: "Métodos Estadísticos en el Ámbito Oficial", dirigido por Gonzalo Mari.

Palabras claves: ESTADÍSTICAS OFICIALES, ANONIMIZACIÓN, INDICADORES DE CALIDAD.

Abstract:

Statistical operations have the objective of collecting data from observation units that can be made up of people, houses, companies or other objects. These data, in the case of official statistics, provide important elements for the evaluation of implemented policies and for future decision-making. Although this is the primary objective, it cannot be ignored that there is a growing trend towards the dissemination of information to society so that private users can analyze the data and be able to obtain their own conclusions. In this way, in recent times, the term open data has had a significant expansion, to the point that there are official portals that contain data that the general public can access to. On the other hand, Law No. 17,622 mentions the obligation to respect statistical secrecy in the disclosure of operations carried out. In this way, there is a contrast between both positions, open data versus statistical secrecy, which is important to take into account to find solutions that allow information to be provided without violating statistical secrecy. It is for this reason that different alternatives that are included within the term anonymization are studied. It is proposed to use perturbative anonymization methods, which modify the observed data of the units to prevent an intruder from detecting, with secondary information, the identity of the unit, which would cause the organization that disseminates the information to violate the mentioned law. Perturbative techniques that add random noise to the data and the method of microaggregation are evaluated. Microaggregation assigns a representative value to all the units that are part of a group of units determined by their proximity. The proposed methods are compared through measurements that account for the distortion produced in the estimates of certain parameters by the application of the perturbative methods to the original data. In this work, certain anonimization scenarios are evaluated on the database of the ENGHO published by the INDEC, using the replicated weight bases for the estimation of certain parameters that quantify the sampling error. In general, the methods that add uncorrelated random noise present better results than their competitors, and more consistent than those produced by the microaggregation method.

Keywords: OFFICIAL STATISTICS, ANONYMISATION, QUALITY INDICATORS.

Introducción

Los operativos estadísticos tienen por objetivo recolectar datos de unidades de observación que pueden estar constituidas por personas, hogares, empresas u otros objetos. Estos datos proporcionan información que, para el caso de las estadísticas oficiales, proporcionan elementos importantes para la evaluación de políticas implementadas y para la toma de decisiones a futuro. Si bien este es el objetivo primario, no puede desconocerse que existe una marcada tendencia en aumento a la difusión de la información a la sociedad para que la misma utilice los datos para analizarlos y poder obtener sus propias conclusiones. De esta forma, en los últimos tiempos, el término datos abiertos tuvo una expansión significativa, al punto tal que existen portales oficiales que contienen datos a los cuales el público en general puede acceder.

Por otra parte, la Ley N° 17.622, en su artículo 10 menciona que

> *Las informaciones que se suministren a los organismos que integran el Sistema Estadístico Nacional, en cumplimiento de la presente ley, serán estrictamente secretos y sólo se utilizarán con fines estadísticos. Los datos deberán ser suministrados y publicados, exclusivamente, en compilaciones de conjunto, de modo que no pueda ser violado el secreto comercial o patrimonial, ni individualizarse las personas o entidades a quienes se refieran. Quedan exceptuados del secreto estadístico los siguientes datos de registro: nombre y apellido, o razón social, domicilio y rama de actividad.*

De esta forma, existe una contraposición entre ambas posturas, datos abiertos versus secreto estadístico, que es importante tener en cuenta para buscar soluciones que permitan brindar información sin violar el secreto estadístico.

En el caso de Argentina, el Instituto Nacional de Estadística y Censos (INDEC) es el organismo responsable de las estadísticas oficiales y coordinador del Sistema Estadístico Nacional (SEN), que está integrado por los Servicios Estadísticos (SE) de los organismos nacionales, provinciales y municipales, que son las unidades orgánicas encargadas de elaborar, recopilar, interpretar y/o divulgar estadísticas oficiales. Respecto al último punto es donde se implementa la ley mencionada, existiendo distintas soluciones de compartir información sin violar el secreto estadístico.

En varios operativos, por ejemplo, los censos de población, hogares y viviendas, la información se comparte a nivel de radios censales, que es una unidad geográfica que agrupa, en promedio, 300 viviendas. En otros, tales como la Encuesta Permanente de Hogares (EPH) o la Encuesta Nacional de Gasto de los Hogares (ENGHO), históricamente se comparten bases usuarios, las cuales son bases de datos con la información recolectada pero que no contiene las variables que determinan el diseño muestral, las cuales son necesarias a la hora de llevar a cabo análisis estadísticos que necesiten realizar inferencias a la población. Una solución que se ha implementado hace unos años es adjuntar a estas bases una nueva con pesos replicados que surgen de un método de replicación, que permite realizar inferencias, pero que debido a la cantidad de réplicas que contiene, las mismas no resultan suficientes para ciertos análisis.

Es por este motivo, que se estudian distintas alternativas que se engloban dentro del término anonimización. En este trabajo, se propone utilizar métodos de anonimización perturbadores, los cuales modifican los datos observados de las unidades para evitar que un intruso pueda detectar, con información secundaria, la identidad de la misma, lo que provocaría que el organismo que difunde la información violase la ley mencionada. En particular, se evalúan las técnicas de perturbación que adicionan un ruido aleatorio a los datos y la de microagregación, que asignan un valor representativo a todas las unidades que forman parte de un grupo de unidades determinado por la cercanía de las mismas.

La solución planteada posee la ventaja de que, al modificar los datos originales, dejaría de existir la posibilidad de emparejar el dato observado con la fuente secundaria de información, y así descubrir la identidad del informante. Pero trae aparejado el inconveniente que los resultados que se obtengan de esta base perturbada, no coincidirán con los que se obtiene de la base original, que es la que utilizan los organismos para publicar sus propias estadísticas. Esto puede levantar un manto de sospecha sobre el proceder de los institutos de estadística y de la información que ellos brindan, que se debería evitar para no perder la credibilidad de los mismos por parte de la sociedad.

Por este motivo, se evalúan los métodos propuestos a través de medidas que dan cuenta de la distorsión que producen en las estimaciones de ciertos parámetros la aplicación de los métodos de perturbación a los datos originales. En el presente trabajo, se realizan pruebas sobre la base usuario de la ENGHO publicada por el INDEC, utilizando las bases de pesos replicados para la estimación de ciertos parámetros que cuantifican el error muestral.

Materiales

Para realizar la evaluación de los métodos de anonimización a través del enfoque de indicadores de evaluación comparativa, se considera la base de datos usuarios de la Encuesta Nacional de Gasto de los Hogares 2017/18 (ENGHO) que realizó el Instituto Nacional de Estadística y Censos (INDEC) entre noviembre de 2017 y noviembre de 2018, y que está disponible en la página web del Instituto[i]. La misma fue realizada en todo el territorio nacional en conjunto con las direcciones provinciales de estadística, y tiene por objetivo principal obtener información acerca de los gastos y los ingresos de los hogares y sus características demográficas (INDEC, 2019). Para este trabajo en particular, se considera solamente la base correspondiente a la provincia de Santa Fe.

Por otra parte, se pueden mencionar objetivos adicionales, pero no menos importantes de esta encuesta, como proporcionar información para el cálculo de las ponderaciones del índice de precios al consumidor (IPC) y actualizar las estructuras de las canastas de bienes y servicios que se emplean para la elaboración de las líneas de pobreza e indigencia.

La ENGHO se realiza en una muestra de hogares que son seleccionados a partir de una muestra probabilística de la Muestra Maestra Urbana de Viviendas de la República Argentina (MMUVRA). Esta muestra maestra considera como población objetivo a aquellas localidades con una población de 2000 y más habitantes de acuerdo al Censo Nacional de Población, Hogares y Viviendas 2010 (Censo 2010) y la misma surge de un muestreo complejo que comprende varias etapas de selección, probabilidades desiguales de selección y distintos niveles de estratificación. Por tal motivo, y sumado al hecho que las encuestas en general están afectadas por la ocurrencia de errores no muestrales tales como errores de cobertura de los marcos y no respuesta, entre otros, la estimación y análisis de los datos recolectados representan un desafío extra debido al hecho que las observaciones no cumplen con los supuestos de independencia e igual distribución, lo que implica que los métodos clásicos inferenciales y de análisis no son válidos para este tipo de encuestas.

Por este motivo, el INDEC comenzó a publicar junto a las bases usuarios, una base adicional con réplicas. Como es sabido, las bases usuario que presenta el INDEC para sus operativos por muestreo no poseen información sobre las variables denominadas de diseño, o sea, las que determinan los estratos de cada una de las etapas, las unidades primarias y secundarias de muestreo, ni las probabilidades de inclusión de cada una de las etapas del diseño. De esta forma, no es posible estimar medidas relacionadas a los errores muestrales, que son de interés en este trabajo, como ejemplos de cambios de estimaciones ante la perturbación de variables.

Para brindar a los usuarios una herramienta que permita realizar análisis de los datos publicados siguiendo el diseño muestral y los ajustes que buscan reducir el sesgo de las estimaciones, se publican bases de pesos replicados los cuales son independientes para cada encuesta y que pueden ser descargados de la página donde está alojada la base usuario correspondiente.

El método de replicaciones que se utiliza para obtener los pesos replicados es el Bootstrap propuesto por Rao y Wu (1988) y en Rao, Wu y Yue (1992) que consiste en definir un número B de submuestras independientes de la muestra original, donde

$B = 200$ para la presente encuesta. En cada una de las submuestras se llevan a cabo los ajustes de pesos que se realizan en los ponderadores de la muestra original con el fin de incorporar la variabilidad que introducen los mismos. Estos ajustes son por no elegibilidad, por no respuesta y calibrados por sexo y edad. Para un detalle más exhaustivo ver INDEC (2020).

El método de Boostrap permite obtener estimaciones válidas de variancia para el estimador de un parámetro θ considerando la variabilidad de las estimaciones del parámetro que se obtienen con cada una de las réplicas. De esta forma se define la variancia Bootstrap de $\hat{\theta}$ a partir de las réplicas como:

$$v_B(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^*_{(b)} - \hat{\theta}\right)^2$$

donde $\hat{\theta}^*_{(b)}$ es el estimador de θ a partir de los ponderadores Bootstrap de la réplica $b = 1, \dots, B$.

Metodología

Métodos de anonimización de datos

Los métodos de anonimización de datos se pueden clasificar en tres tipos:

- Técnicas no perturbadoras, como la recodificación o eliminación local, que reducen o suprimen el nivel de detalle sin alterar los datos originales;
- Técnicas perturbadoras, como la adición de ruido, método de post-aleatorización, microagregación;
- Técnicas que generan bases de microdatos sintéticas que preservan ciertas estadísticas o relación entre las variables originales.

En este trabajo se aplicaron dos técnicas perturbadoras aplicadas a variables continuas: ruido aditivo (no correlacionado y correlacionado) y microagregación. En primer lugar, se consideró el ruido aditivo no correlacionado. Sea $\boldsymbol{X}$ la matriz de los valores originales para las p variables donde:

$$\boldsymbol{X} \sim (\mu, \sigma)$$

Se definen las variables perturbadas como:

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

donde $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$ y $\boldsymbol{\Sigma}_\epsilon = c\ diag\left(\sigma_1^2, \dots, \sigma_p^2\right)$ para una constante $c > 0$.

En segundo lugar, se consideró el ruido aditivo correlacionado. Este método es similar al anterior, pero difiere en la forma en que se define la matriz de covariancias del ruido adicionado, siendo la misma:

$\boldsymbol{\Sigma}_\epsilon = c\boldsymbol{\Sigma}$ lo que implica que $\boldsymbol{\Sigma}_{\boldsymbol{Z}} = \boldsymbol{\Sigma} + c\boldsymbol{\Sigma} = (1 + c)\boldsymbol{\Sigma}$.

El segundo método considerado fue el de microagregación, que es un método perturbativo que particiona los registros en grupos de características similares, y reemplaza los valores de las variables por una medida resumen (como por ejemplo la media aritmética o la mediana). Los grupos pueden formarse a partir de distintos criterios como el método de los rankings individuales, el basado en análisis de componentes principales y el método de Máxima Distancia al Valor Medio (MDAV) que es el utilizado en este caso.

Se plantearon 6 escenarios de anonimización en base al riesgo de divulgación (5%,10%,25%,50%,75%). El criterio utilizado para este riesgo fue el correspondiente al de divulgación por intervalo, que se calcula en base al porcentaje de observaciones originales que caen dentro del intervalo calculado sobre las observaciones perturbadas en base a su variabilidad. Para aplicar los 3 métodos de anonimización se utilizó el paquete sdcMicro del programa R. Se aplicaron los tres métodos de anonimización en cada uno de los escenarios planteados, modificando algunos parámetros para obtener el porcentaje de riesgo de divulgación deseado en cada caso. Esto se realizó de la siguiente manera:

- Riesgo aditivo no correlacionado: Se adicionó el ruido no correlacionado (*method="additive"*) modificando el parámetro de ruido hasta lograr el riesgo de divulgación propuesto en el escenario considerado. Los restantes parámetros fueron los propuestos por defecto por el paquete sdcMicro.
- Riesgo aditivo correlacionado: Se adicionó el ruido correlacionado (*method="correlated"*) modificando el parámetro de ruido hasta lograr el riesgo de divulgación propuesto en el escenario considerado. Los restantes parámetros fueron los propuestos por defecto por el paquete sdcMicro.
- Microagregación: se realizó la microagregación modificando el parámetro correspondiente al nivel de agregación (cantidad de observaciones en cada grupo) hasta lograr el riesgo de divulgación propuesto en el escenario considerado. El método para crear los grupos fue el de Máxima Distancia al Valor Medio (*method="mdav"*). Los restantes parámetros fueron los propuestos por defecto por el paquete sdcMicro.

Utilidad de los datos y pérdida de información

Una vez aplicados los métodos de anonimización para modificar el conjunto de datos original y para reducir el riesgo de divulgación, es fundamental medir la pérdida de información resultante y la utilidad de los datos. Básicamente, existen dos tipos diferentes de enfoques para evaluar la pérdida de información:

- medición directa de las distancias/frecuencias entre los datos originales y los perturbados, y
- comparación de las estadísticas calculadas con los datos originales y los perturbados.

El primer concepto es común, pero a menudo de uso limitado mientras que el segundo es de mayor utilidad para los usuarios dado que tendrá consecuencias directas en las estimaciones que estos calculen.

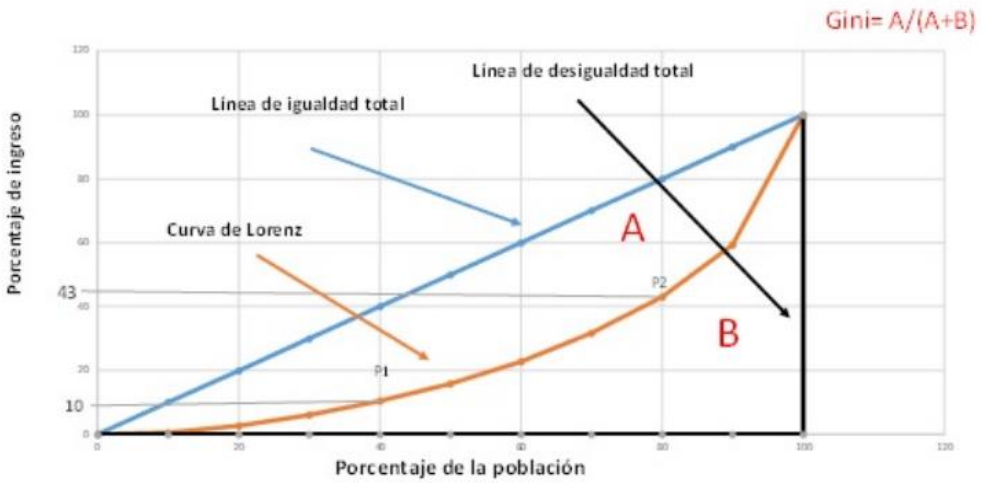
Se consideró una medida para medir la calidad de manera tal que la técnica de anonimización minimice las diferencias entre las propiedades estadísticas de los datos perturbados en comparación con los datos originales. El objetivo es que aquellas estimaciones consideradas importantes no presenten cambios significativos al ser computadas con los datos perturbados. A tal fin se seleccionó el Coeficiente de Gini como un estimador de relevancia, calculado con los datos originales y los perturbados. El Coeficiente de Gini es una medida bien conocida de desigualdad de una distribución y es ampliamente aplicado en muchos campos de investigación. Dado un vector $x_1, x_2, \dots, x_n$con pesos muestrales $w_1, w_2, \dots, w_n$, el coeficiente se estima a partir de:

$$\hat{G} = 100\left[\frac{2\sum_{i=1}^{n}\left(w_i x_i \sum_{j=1}^{n} w_j\right) - \sum_{i=1}^{n} w_i^2 x_i}{\left(\sum_{i=1}^{n} w_i\right)\sum_{i=1}^{n} w_i x_i} - 1\right]$$

El Coeficiente de Gini es una medida de desigualdad que generalmente se utiliza para variables relacionadas con los ingresos o gastos de las personas u hogares, que toma valores entre 0 y 100. Un valor de 0 significa una perfecta igualdad, lo que implica que cada unidad en la muestra tiene el mismo valor o todos los indivíduos el mismo valor de la variable. Con un valor de 100 el coeficiente de Gini indicaría una absoluta desigualdad, lo que significa que todos los datos menos uno, son iguales a cero o que solo un individuo tiene un valor positivo de la variable, por ejemplo, acumula todo el ingreso o volumen de un cierto bien. El Coeficiente de Gini se lo relaciona con la Curva de Lorentz, como se puede apreciar en el gráfico siguiente, donde el coeficiente coincide con el cociente entre el área A y la suma de las áreas A y B multiplicado por 100, es decir:

$$G = \frac{A}{(A+B)}100$$

Gráfico 1: Curva de Lorenz y Coeficiente de Gini



Para evaluar los cambios presentados se consideraron las estimaciones puntuales, las estimaciones de variancias y los porcentajes de cobertura de los intervalos de confianza para este índice.

En el caso referido a los cambios en las estimaciones puntuales y de las variancias, esto se realizó de acuerdo con la siguiente fórmula:

$$\frac{\left|\hat{\theta}_{pert} - \hat{\theta}_{orig}\right|}{\hat{\theta}_{orig}} \times 100$$

donde $\hat{\theta}$ es la estimación del Coeficiente de Gini o la estimación de su variancia. Los subíndices indican si la estimación se realizó con los datos originales (*orig*) o con los datos perturbados (*pert*). La estimación de las variancias se realizó utilizando el

estimador de variancia Bootstrap de $\hat{\theta}$ a partir de las réplicas de la muestra, como se explica en la sección Materiales.

Los intervalos de confianza se calcularon utilizando las estimaciones puntuales y sus correspondientes estimaciones de variancia Bootstrap, tanto para los datos originales como para los perturbados, asumiendo distribución normal, de la siguiente manera:

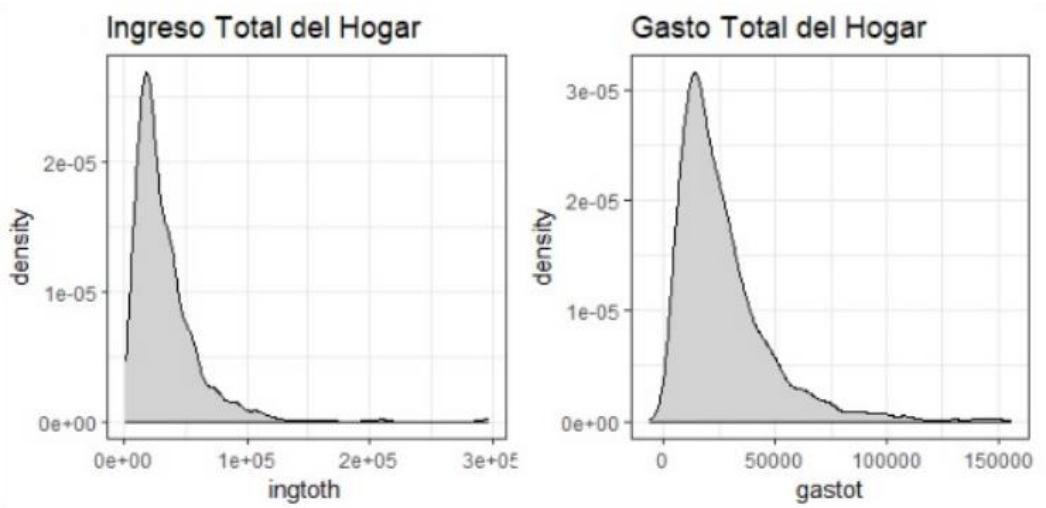
$$\hat{\theta} \pm 1{,}96\sqrt{v_B(\hat{\theta})}$$

En el caso de los porcentajes de cobertura de los intervalos de confianza, estos se computaron de manera tal que cada valor obtenido indica qué porcentaje del intervalo de confianza del Coeficientes de Gini para la variable perturbada se superpone con el intervalo de confianza obtenido para la variable sin modificar, en cada uno de los escenarios propuestos.

Resultados

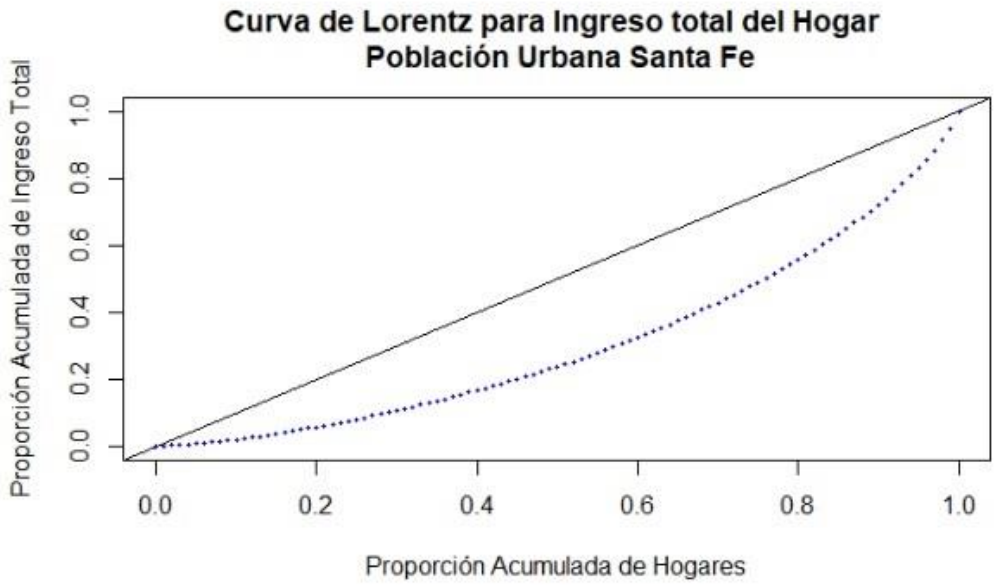
Se analizaron los datos provenientes de 1.122 hogares de la Provincia de Santa Fe. En el Gráfico 2 se presentan las distribuciones de las variables Ingreso Total del Hogar (ingtoth) y Gasto Total del Hogar (gastot), las cuales son marcadamente asimétricas a la derecha.

Gráfico 2: Distribuciones de las vables Ingreso Total y Gasto Total del Hogar



En el Gráfico 3 se puede observar la curva de Lorentz para el Ingreso Total del Hogar (ingtoth), la cual es similar a la del Gasto Total. Si el ingreso fuese similar en todos los hogares, la curva de puntos coincidiría con la recta de trazo sólido y el coeficiente de Gini tomaría su mínimo valor igual a 0. Si existiese una situación de total desigualdad donde un solo hogar concentre el total de los ingresos y el resto de los hogares tuviesen ingresos nulos, el coeficiente de Gini tomaría el valor 100.

Gráfico 3



Las estimaciones del Coeficiente de Gini para ambas variables se muestra en la Tabla 1. Las variancias de las estimaciones se computaron con las réplicas bootstrap y los intervalos de confianza asumen distribución normal para el estimador.

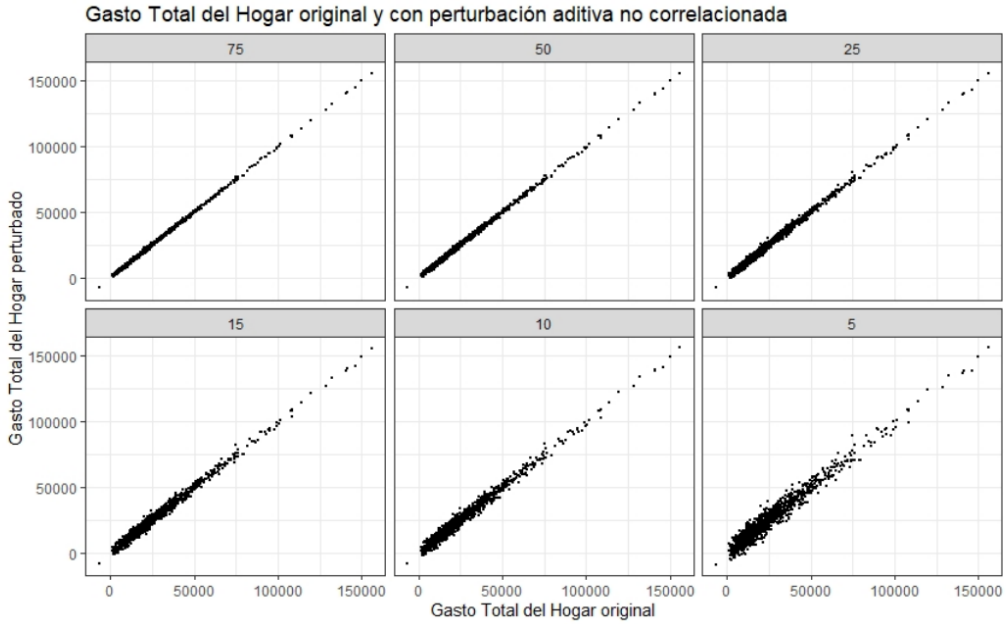
Tabla 1: Estimaciones puntuales y por intervalo para el coeficiente de Gini para Ingreso y Gasto Total del hogar

Variable	$\hat{G}$	$\hat{S}(\hat{G})$	l.i. del IC95 %	l.s. del IC95 %
ingtoth	38.34	1.03	36.46	40.51
gastot	38.64	1.11	36.64	41.00

Se presentan los resultados de los 6 escenarios para los 3 métodos de anonimización analizados: ruido aditivo no correlacionado, correlacionado y microagregación. Estos escenarios presentan distintos riesgos de divulgación: 75 %, 50 %, 25 %, 15 %, 10 %, y 5 %. Es decir, se aplicaron los métodos perturbando los datos con el fin de ir reduciendo dicho riesgo.

En el Gráfico 4 se muestra como se modifica la variable Gasto Total al aplicar el método de ruido aditivo no correlacionado. Se puede observar que a medida que se agrega mayor ruido y se reduce el riesgo de divulgación, la nube de puntos en el diagrama de dispersión que muestra los valores originales y perturbados se vuelve más dispersa. Los datos iniciales quedan más alejados de los modificados y por ende se reduce el riesgo de divulgación. Este resultado es similar para Ingreso Total.

Gráfico 4



De manera similar, en el Gráfico 5 se muestra cómo se modifica la variable Gasto Total al aplicar el método de ruido aditivo correlacionado. Se puede observar que a medida que se agrega mayor ruido y se reduce el riesgo de divulgación, la nube de puntos en el diagrama de dispersión que muestra los valores originales y perturbados también se vuelve más dispersa. Lo mismo ocurre para Ingreso Total.

Gráfico 5

El Gráfico 6 muestra como se modifica la variable Gasto Total al aplicar el método de microagregación. Los resultados de la perturbación en este método difieren de los dos

anteriores por la naturaleza del mismo. En este caso, a medida que se reduce el riesgo de divulgación, el método necesita crear grupos más numerosos de observaciones cercanas a las cuales se les remplaza su valor por un valor común a ellas. Si bien se logra mejorar la anonimización, se crean grupos de observaciones con el mismo valor, lo cual modifica en cierta manera el carácter continuo de la variable y reduce su variabilidad.

Gráfico 6

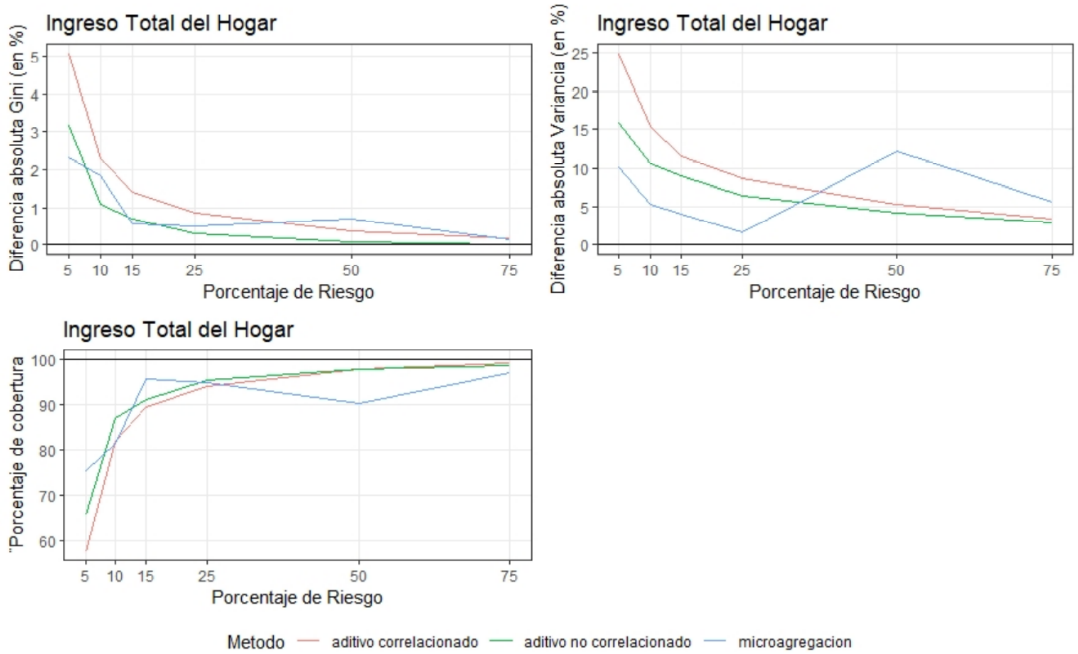
Se presentan los resultados de las diferencias absolutas porcentuales de las estimaciones con las variables originales y las variables modificadas en cada uno de los escenarios. Esto se realizó tanto para el coeficiente de Gini como para su variancia. También se presentan los resultados para los porcentajes de cobertura de los intervalos de confianza.

En el Gráfico 7 se pueden ver los resultados para Ingreso Total del Hogar. En cuanto a las diferencias absolutas en la estimación del coeficiente de Gini se observa que estas en general son mayores para situaciones donde el riesgo de divulgación es menor y disminuyen cuando los riesgos son mayores. Si se considera un riesgo de divulgación del 5%, las diferencias oscilan aproximadamente entre un 2 y un 5%. Para un riesgo del 15% las diferencias son menores al 2% y para riesgos del 25% o mayores todas las diferencias son inferiores al 1%. Los métodos que adicionan ruido muestran una reducción consistente de las diferencias a medida que aumenta el riesgo, es decir que siempre las diferencias disminuyen a mayor riesgo. En el caso del método de microagregación, si bien las diferencias disminuyen para los riesgos más altos, existen algunos casos donde no necesariamente lo hacen, siendo estos los correspondientes al riesgo de 25 y 50%. Aun así, este método muestra todas diferencias absolutas menores al 1% a partir de riesgos del 15% o mayores, con lo cual este comportamiento no parece ser relevante. Pareciera que para riesgos de divulgación bajos (del 5 al 15%) tanto ruido no correlacionado como microagregación tuvieran los mejores desempeños siendo el primero levemente superior para riesgos más altos.

Con respecto a las diferencias absolutas de las estimaciones de las variancias, si se considera un riesgo de divulgación del 5%, las diferencias oscilan aproximadamente entre un 10 y un 25%. Para un riesgo del 15% las diferencias son menores al 15% y para un riesgo del 25% todas las diferencias son menores al 10%. Los métodos que adicionan ruido muestran mayores diferencias a medida que se reduce el riesgo de divulgación. Esto no se cumple para microagregación, que muestra un valor más elevado para un riesgo del 50%. En el caso de riesgos de divulgación más bajos (del 5 al 25%) microagregación presenta un mejor desempeño siendo el ruido aditivo correlacionado el de peores resultados. Para riesgos más elevados el ruido aditivo no correlacionado brinda las menores diferencias.

Con respecto a los porcentajes de cobertura, se observa que estos en general son menores para situaciones donde el riesgo de divulgación es menor y aumentan cuando los riesgos son mayores. Si se considera un riesgo de divulgación del 5%, los porcentajes oscilan aproximadamente entre un valor levemente menor al 60% y un valor levemente superior al 75%. Para un riesgo del 10% los porcentajes oscilan entre 80 y 90% y para riesgos del 15% o mayores todos los porcentajes de cobertura son superiores al 90% excepto un caso que está levemente por debajo del 90%. Los métodos que adicionan ruido muestran una reducción consistente de los porcentajes a medida que disminuye el riesgo, es decir que siempre los porcentajes de cobertura disminuyen a menor riesgo. En el caso del método de microagregación, si bien los porcentajes aumentan para los riesgos más altos, existen algunos casos donde no necesariamente aumentan, siendo estos los correspondientes al riesgo de 25 y 50%. Aun así, este método presenta porcentajes de cobertura con valores todos superiores al 90% a partir de riesgos del 15% o mayores, con lo cual este comportamiento no resulta importante. Pareciera que para riesgos de divulgación bajos (del 5 al 15%) tanto ruido no correlacionado como microagregación tuvieran los mejores desempeños, presentando el primero resultados mejores para riesgos superiores.

Gráfico 7: Ingreso Total del Hogar. Diferencias absolutas en las estimaciones del Coeficiente de Gini y sus variancias. Porcentajes de cobertura de los intervalos de confianza.



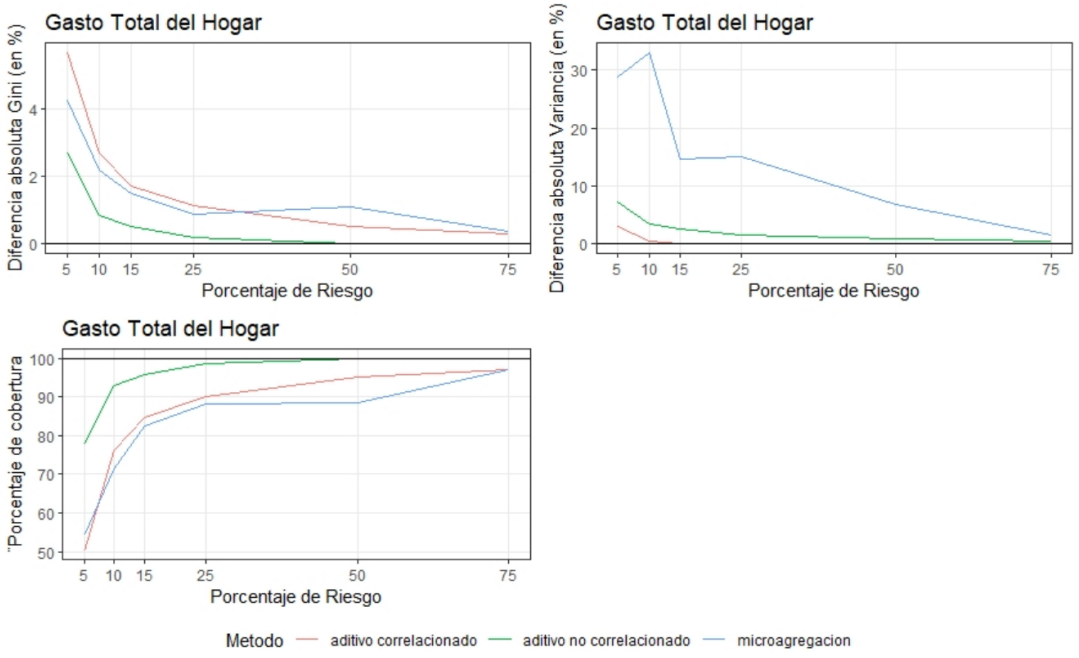
En el Gráfico 8 se pueden ver los resultados para Gasto Total del Hogar. En cuanto a las diferencias absolutas en la estimación del coeficiente de Gini se observa, de manera similar que para el Ingreso Total, que las mismas en general son mayores para situaciones donde el riesgo de divulgación es menor y disminuyen cuando los riesgos son mayores. Si se considera un riesgo de divulgación del 5%, las diferencias oscilan aproximadamente entre un 2 y un 6%. Para un riesgo del 15% o mayor las diferencias son menores al 2%. Los métodos que adicionan ruido muestran una reducción consistente de las diferencias a medida que aumenta el riesgo, es decir que siempre las diferencias disminuyen a mayor riesgo, como en el caso de Ingreso Total. En el caso del método de microagregación, si bien en general las diferencias disminuyen para los riesgos más altos, esto no se cumple para el riesgo de 50%. De todas formas este valor es levemente superior al correspondiente a un riesgo del 25% Se puede observar que para todos los riesgos de divulgación planteados, el ruido aditivo no correlacionado tiene el mejor desempeño en este caso.

Con respecto a las diferencias absolutas de las estimaciones de las variancias, si se considera un riesgo de divulgación del 5%, las diferencias oscilan aproximadamente entre un 10 y un 25%. Para un riesgo del 15% las diferencias son menores al 15% y para un riesgo del 25% todas las diferencias son menores al 10%. Los métodos que adicionan ruido muestran mayores diferencias a medida que se reduce el riesgo de divulgación. Esto no se cumple para microagregación, que muestra un valor más elevado para un riesgo del 50%. En el caso de riesgos de divulgación más bajos (del 5 al 25%) microagregación presenta un mejor desempeño siendo el ruido aditivo correlacionado el de peores resultados. Para riesgos más elevados el ruido aditivo no correlacionado brinda las menores diferencias.

Con respecto a las diferencias absolutas de las estimaciones de las variancias, si se considera un riesgo de divulgación del 5%, las diferencias parecen ser importantes entre los métodos, ya que los que adicionan ruido presentan valores menores al 10% y micoragregación se ubica levemente por debajo del 30%. Estas diferencias se repiten para riesgos superiores (10 al 25%). Los métodos que adicionan ruido muestran mayores diferencias a medida que se reduce el riesgo de divulgación. Esto no se cumple para microagregación, que muestra valores más elevados para los riesgos del 10% y el 25%. El método de ruido aditivo correlacionado presenta el mejor desempeño en este caso, mientras que el ruido no correlacionado tiene un comportamiento cercano, con valores levemente más altos.

Con respecto a los porcentajes de cobertura, se observa que estos son menores para situaciones donde el riesgo de divulgación es menor y aumentan cuando los riesgos son mayores. Si se considera un riesgo de divulgación del 5%, se observa una diferencia importante entre el ruido no correlacionado con un valor levemente por debajo del 80% y los otros dos métodos que presentan valores por debajo del 60%. Estas diferencias, aunque menores, se mantienen para riesgos de divulgación más altos (del 10% al 25%). El ruido aditivo no correlacionado tiene mejor desempeño en este caso. Los otros dos métodos parecen tener un desempeño similar con una cierta ventaja a favor del ruido aditivo correlacionado.

Gráfico 8: Gasto Total del Hogar. Diferencias absolutas en las estimaciones del Coeficiente de Gini y sus variancias. Porcentajes de cobertura de los intervalos de confianza.



Conclusiones

Al evaluar los resultados considerando los escenarios planteados se puede concluir que para la variable Ingreso Total del Hogar los métodos de ruido aditivo no correlacionado y microagregación presentan los mejores desempeños con una cierta ventaja para el primero. En el caso de la variable Gasto Total del Hogar se puede afirmar que el ruido aditivo no correlacionado presenta los mejores resultados, basados en las medidas de calidad expuestas.

Los métodos que adicionan ruido, ya sea correlacionado o no, presentan resultados intuitivamente más consistentes con respecto a los porcentajes de riesgo de divulgación planteados. Si bien el método de microagregación presenta un desempeño aceptable para Ingreso Total del Hogar, muestra una inestabilidad que no tiene una explicación lógica y que puede deberse a la naturaleza de la distribución de las variables. El método genera varios valores iguales para distintas observaciones, limitando el origen continuo de la variable y disminuyendo su variabilidad. Esta pérdida de información puede perjudicar o inhabilitar ciertos análisis de datos para esas variables.

Los resultados obtenidos pueden no coincidir con resultados obtenidos previamente. En la bibliografía se pueden encontrar trabajos que comparan los efectos sobre los estimadores de los distintos métodos de anonimización pero sin controlar el riesgo de divulgación. Dado que el objetivo principal es proteger la identidad de las unidades que brindan información, es importante tener en cuenta este riesgo a la hora de comparar los métodos. En principio pareciera que los análisis y los resultados van a depender de

las variables definidas como claves o esenciales, el porcentaje de riesgo de divulgación dispuesto a admitir y los indicadores de calidad especificados.

En el futuro se continuará trabajando con otras variables que incluyan además a variables categóricas y otros métodos de anonimización.

Referencias Bibliográficas

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K.,de Wolf, P-P. (2012). *Statistical disclosure control*. Wiley Series in Survey Methodology: Wiley. ISBN 9781118348222.

INDEC (2018). Estudio nacional sobre el perfil de las personas con discapacidad: resultados provisorios 2018. - 1a ed. Ciudad Autónoma de Buenos Aires: Instituto Nacional de Estadística y Censos - INDEC. Libro digital, PDF.

INDEC (2019). Encuesta Nacional de Gastos de los Hogares 2017-2018 : resultados preliminares / 1a ed. Ciudad Autónoma de Buenos Aires: Instituto Nacional de Estadística y Censos - INDEC. Libro digital, PDF. Recuperado el 8/5/2021 de https://www.indec.gob.ar/ftp/cuadros/sociedad/engho_2017_2018_resultados_preliminares.pdf

INDEC (2020). Encuesta Nacional de Gastos de los Hogares 2017-2018 : Factores de expansión, estimación y cálculo de los errores de muestreo : nota técnica n°4. - 1a ed . - Ciudad Autónoma de Buenos Aires : Instituto Nacional de Estadística y Censos – INDEC. Libro digital, PDF. Recuperado el 8/5/2021 de https://www.indec.gob.ar/ftp/cuadros/menusuperior/engho/engho2017_18_nota_tecnica_4.pdf

Rao, J. N. K. y Wu, C. F. J. (1988). Resampling Inference with Complex Surveys Data. *Journal of American Statistical Association*, 83, 231-241. DOI: 10.1080/01621459.1988.10478591.

Rao, J. N. K., Wu, C. F. J. y Yue, K. (1992). Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, 18, 209-217. Recuperado el 8/5/2021 de: https://www150.statcan.gc.ca/n1/pub/12-001-x/1992002/article/14486-eng.pdf.

Templ, M. (2017). *Statistical Disclosure Control for Microdata*. Springer International Publishing: Basel, Switzerland. ISBN 978-3-319-50270-0.

Templ, M., Kowarik, A., Meindl B. (2015). Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. Journal of Statistical Software, 67(4), 1-36.<doi:10.18637/jss.v067.i04>

[i] https://www.indec.gob.ar/indec/web/Institucional-Indec-BasesDeDatos-4