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
Marí, Gonzalo
Méndez, Fernanda
Instituto de investigación, Escuela de Estadística

MÉTODOS DE ANONIMIZACIÓN DE BASES DE DATOS EN ESTADÍSTICAS OFICIALES[1]

Resumen

En Argentina, la Ley Nº 17.622 del año 1968, establece que la información que se suministre a los organismos integrantes del Sistema Estadístico Nacional está protegida por el secreto estadístico, el cual asegura que los datos deberán ser publicados de forma tal que no pueda individualizarse a las unidades a quienes se refieran los mismos. Es por este motivo, que los organismos oficiales de estadística establecen una serie de medidas tendientes a asegurar el anonimato de los informantes. Estas acciones, en muchos operativos, se reducen a la no publicación de las bases de datos, siendo las estadísticas que se publican la única información disponible para los mismos. De esta forma, no existe la posibilidad que los usuarios desarrollen sus propios análisis sobre los conjuntos de datos.

Otra situación se presenta en el caso de las encuestas por muestreo, donde una práctica habitual es la difusión de bases de datos de usuarios, las cuales contienen las variables que fueron medidas durante el operativo, pero carecen de aquellas referidas al diseño muestral, que permitiría, en un hipotético caso, identificar a las unidades a las cuales se refiere la información. Esta decisión posee la desventaja de que sin esa información no es posible llevar a cabo análisis adicionales apropiados sobre las bases de datos debido a que no es posible considerar las características del diseño muestral empleado, que, en la mayoría de los casos, es complejo y no emplear la información del mismo en el análisis de los datos conlleva a que los resultados no sean válidos.

En ambas situaciones la anonimización de las unidades de análisis está garantizada, si bien la pérdida de información es demasiado importante, en un caso por la ausencia total de la misma y en el segundo caso por la imposibilidad de poder emplear el diseño muestral en los análisis posteriores. De esta forma, debe existir un balance entre la pérdida o distorsión de la información que se publica y la anonimización de las unidades que brindaron los datos, de tal forma que no exista riesgo de divulgación de la identidad, pero con una mínima perdida de información que permita que la base de datos protegida continúe siendo útil.

En el presente trabajo se estudian distintos métodos de anonimización de bases de datos, tanto para variables categóricas como cuantitativas, y se evalúan los mismos considerando medidas relacionadas con el riesgo de divulgación y la pérdida de información que ocasionan las metodologías consideradas.

Palabras claves: estadísticas oficiales, riesgo de divulgación, anonimización.

[1] Trabajo elaborado en el marco del Proyecto ECO 1ECO199 titulado “Métodos Estadísticos en el Ámbito Oficial”, dirigido por Gonzalo Marí

Abstract

In Argentina, Law No. 17.622 of the year 1968, establishes that the information provided to the member organizations of the National Statistical System is protected by statistical secrecy, which ensures that the data must be published in such a way that no individuals or entities to which they refer to can be individualized. It is for this reason that the official statistical agencies establish a series of actions aimed at ensuring the anonymity of informants. These actions, in many operations, are reduced to the non-publication of the databases, with the statistics published being the only information available to them. In this way, there is no possibility for users to develop their own statistical analysis of data sets.

Another situation arises in the case of sampling surveys, where a common practice is the dissemination of user databases, which contain the variables that were measured during the operation, but lack those referred to the sample design, which would allow, in a hypothetical case, to identify the units to which the information refers to. This decision has the disadvantage that without this information it is not possible to carry out additional proper analyzes on the databases because it is not possible to consider the characteristics of the sample design used, which, in most cases, is complex and not using this information in the analysis of the data means that the results would not be valid.

In both situations, the anonymization of the analysis units is guaranteed, although the loss of information is extremely important, in one case due to the total absence of it and in the second case due to the impossibility of being able to use the sample design in the analyzes. Thus, there must be a balance between the loss or distortion of the information that is published and the anonymization of the units that provided the data, so that there is no risk of disclosure of identity, but with a minimum loss of information that allows the protected database to continue to be useful.

In this paper we study different methods of anonymization of databases, both for categorical and quantitative variables, and they are evaluated considering measures related to the risk of disclosure and the loss of information caused by the methodologies considered.

Keywords: official statistics, disclosure risk, anonimization.

Objetivos

Se plantea como objetivos del presente trabajo, estudiar y aplicar métodos que permitan la anonimización de bases de datos provenientes de encuestas por muestreo, considerando variables categóricas y cuantitativas.

Metodología y análisis de datos considerados en la investigación

Se considera la aplicación de distintas técnicas determinísticas y probabilísticas de anonimización de bases de datos provenientes de encuestas por muestreo. En el primer grupo, la recodificación es un método no perturbativo que puede ser aplicado tanto a variables categóricas como continuas. Para las primeras, el método consiste en combinar categorías de la variable en nuevas categorías menos informativas, mientras que, para variables continuas, se considera la discretización de la misma. Otro método determinístico es el de eliminación local, que también es no perturbativo y que se aplica generalmente a variables categóricas. Consiste en eliminar valores individuales de unidades con alto riesgo individual de descubrimiento, de forma tal de alcanzar anonimidad de acuerdo a un cierto criterio.

Entre los métodos probabilísticos, se puede mencionar el Método de Post-Aleatorización (PRAM), un método perturbativo que puede aplicarse a variables categóricas, recodificando los

valores en otras categorías teniendo en cuenta probabilidades de transición conocidas.

Con respecto a los métodos para variables cuantitativas se menciona la micro-agregación, 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 (media aritmética, mediana, etc). Los grupos pueden formarse a partir de diversas herramientas como el ranking de una variable, componentes principales, distancia euclídea, distancia robusta de Mahalanobis, etc.

Otra opción consiste en agregar ruido a variables continuas, el cual es un método perturbativo. El método consiste en sumar o multiplicar los valores originales por un valor estocástico o aleatorio. Pueden ser considerados distintos algoritmos, entre los cuales se destacan aquellos que suman ruido aditivo no correlacionado, ruido aditivo correlacionado, y ruido correlacionado basado en transformaciones, entre otros.

Se aplican las técnicas de anonimización a bases de datos de encuestas desarrolladas por el INDEC. Para los métodos correspondientes a variables categóricas, se considera una base preliminar correspondiente Estudio Nacional sobre el Perfil de las Personas con Discapacidad realizada durante el año 2018, realizado a partir de un convenio entre el INDEC y la Agencia Nacional de Discapacidad (INDEC, 2018). En el caso de los métodos correspondientes a variables cuantitativas, se utiliza una base de datos preliminar correspondiente a la Encuesta Nacional de Gasto de los Hogares llevada a cabo por INDEC en 2017/18. Se evalúa cada uno de los métodos a través de medidas relacionadas con la anonimidad de las bases, el riesgo de divulgación de las unidades que conforman las mismas, y con la pérdida de información.

Problemas planteados, principales hipótesis y resultados

Se consideran diversos métodos de anonimización de bases de datos con el objetivo de evaluar los mismos y plantear distintas alternativas a la publicación de bases de datos usuarias por parte de Oficinas Nacionales de Estadística. Hasta la actualidad no existen estudios a partir de los cuales se obtenga una medición de cuál es el riesgo de violar el secreto estadístico a partir de la divulgación de información de las unidades que conforman los operativos estadísticos. Se considera necesario realizar este tipo de análisis que permite evaluar los riesgos y proponer métodos que minimicen los mismos sin la consabida pérdida de información.

Descripción de la novedad y relevancia del trabajo

No existe hasta la actualidad estudios empíricos referidos a la anonimización de bases de datos en estadísticas oficiales en el país. El presente trabajo constituye un punto de partida a través del cual será posible obtener una cuantificación del riesgo de violar el secreto estadístico a la hora de divulgar bases de datos, y que brinda algunos elementos necesarios para solucionar estos problemas intentando minimizar la pérdida de información.

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.

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>