UNECE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Confidentiality

1-3 December 2021, Poland

Proposal for a risk assessment scale for privacy risks in the disclosure of statistical information.

Jesús González López (INEGI)
jesus.gonzalez@inegi.org.mx

Abstract

The United Nations Fundamental Principles of Official Statistics (UNFPOS) establish in the principle 6 that Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes. The Implementation Guidelines for the United Nations Fundamental Principles of Official Statistics establish in Part I Confidentiality, that data producers must carry out activities to avoid direct and indirect disclosure of individual data. Finally, the United Nations National Quality Assurance Framework Manual for Official Statistics (UNNQAF MANUAL) establishes to ensure statistical confidentiality it is necessary to assess the identification risk and maintain a balance between the acceptable level of identification risk of the individual respondents and the data usability.

Each of the previous references, recommends evaluating the risk of direct and indirect identification, but how to do it? It is understood that the level of identification risk varies according to the characteristics of the information, its measurement can be determined in a metric such as high, medium, or low. The point is how to measure it? And what to do after the measurement? This article presents a proposal based on the hypothesis that a qualitative assessment scale facilitates the analysis and treatment of identification risks. The scale is designed to be applied before the information is disseminated to decide whether the information can be disseminated in the terms in which it is or it needs to be disseminated differently, to maintain a balance between privacy and accuracy of information. Although the scale was conceived for Mexico's national statistical office, it is not ruled out that it is useful for other similar ones.

Proposal for a risk assessment scale for privacy risks in the disclosure of statistical information.

Luis Martín Clemente Aréchiga, INEGI, Luis.Clemente@inegi.org.mx
Jesús González López, INEGI, Jesus.Gonzalez@inegi.org.mx

Abstract: The United Nations Fundamental Principles of Official Statistics (UNFPOS) establish in the principle 6 that Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes. The Implementation Guidelines for the United Nations Fundamental Principles of Official Statistics establish in Part I Confidentiality, that data producers must carry out activities to avoid direct and indirect disclosure of individual data. Finally, the United Nations National Quality Assurance Framework Manual for Official Statistics (UNNQAF MANUAL) establishes to ensure statistical confidentiality it is necessary to assess the identification risk and maintain a balance between the acceptable level of identification risk of the individual respondents and the data usability.

Each of the previous references, recommends evaluating the risk of direct and indirect identification, but how to do it? It is understood that the level of identification risk varies according to the characteristics of the information, its measurement can be determined in a metric such as high, medium, or low. The point is how to measure it? And what to do after the measurement? This article presents a proposal based on the hypothesis that a qualitative assessment scale facilitates the analysis and treatment of identification risks. The scale is designed to be applied before the information is disseminated to decide whether the information can be disseminated in the terms in which it is or it needs to be disseminated differently, to maintain a balance between privacy and accuracy of information. Although the scale was conceived for Mexico's national statistical office, it is not ruled out that it is useful for other similar ones.

1 The need to measure and know

The State, as a political organization responsible for the administration of public affairs in a country, needs to know the number and characteristics of the population as well as the activities carried out in the territory it governs. Since ancient times, statistics have allowed rulers to obtain these data, which at the time have facilitated decision-making and government actions. According to the historian Herodotus, in Egypt around 3050 BC specific data on the country's population and wealth was collected as part of the preparations prior to the construction of the pyramids (Hernández, 2005). Other important antecedents are the Census of China in 2238 BC and the population counts in Babylon and Rome (Quintela, 2019). In each of the previous examples, the data generated through statistics allowed governments to measure and learn to make decisions regarding the administration, direction, and control of the governed territory.

The data generated from statistical exercises such as censuses and surveys allow governments to make decisions to plan, measure and correct the economic and social development of each nation. This type of information is usually called official statistics whose data reveal different needs, improvements or setbacks in different areas and phenomena in each period. It is increasingly necessary for official statistics to provide granular data, especially in democratic systems where official statistics are expected to inform about minorities and vulnerable groups so they may be considered and favored in the design of public policies to guarantee them access to public services and the exercise of fundamental rights. For example, suppose that a government has the objective of improving the conditions of ethnic groups. To achieve this, it is necessary to know at least how many ethnic groups exist, how many members its community is made up of and most likely, in what part of the territory are established. Based on the above data, the government will be able to identify their needs and formulate actions to improve their wellbeing.

2 The commitment to inform without exposing identities

The fundamental input to produce official statistics is the data provided by people and organizations (data providers). The quality and precision of the information generated is directly proportional to the veracity of their answers. The statistical laws of several countries establish that people and organizations are obliged to provide data to statistical agencies. The same laws also "protect the data provider against any interference, protecting it from disclosures that may be annoying or cause moral or moral damage. economic, to encourage him not to hide or distort the information” (Masciadri, 2013, p.144). Therefore, it can be distinguished that statistical agencies have two major responsibilities, on the one hand to produce quality information for decision-making and, on the other hand, to maintain statistical confidentiality, which consists of “the prohibition of disseminating data in those who do not preserve the anonymity of each individual unit to which the information refers” (Masciadri 2013 p.144).

The Fundamental Principles of Official Statistics establish in Principle 6 that "the individual data collected by statistical agencies for statistical compilation refer to natural or legal persons, must be strictly confidential." For its part, UN Fundamental Principles of Official Statistics - Implementation guidelines, establishes that to preserve the confidentiality of data providers, statistical agencies must apply statistical disclosure controls before the information is disseminated, as well as carry out a review of the information before it is disseminated to ensure that the anonymity of the data providers is not compromised. Additionally, the United Nations National Quality Assurance Frameworks Manual for Official Statistics (UN NQAF Manual) recommends that to preserve the confidentiality of data providers the identification risk should be assessed and documented.

Regardless of whether statistical agencies have the responsibility of implementing mechanisms to preserve statistical confidentiality throughout the information life cycle, there is a set of activities that must be carried out prior to the dissemination of information to maintain the statistical confidentiality:

Reference	Requirements
UN Fundamental Principles of Official Statistics – Implementation guidelines	Statistical data producers apply statistical disclosure control methods prior to the release of statistical information. Review by authorized staff of all data prepared for dissemination for possible indirect disclosure.
United Nations National Quality Assurance Frameworks Manual for Official Statistics (UN NQAF Manual)	There should be a balance between the acceptable level of risk of identification of individual respondents and usability of the data. Appropriate processes are in place to assess the risk of disclosure of sensitive information

3 Proposed scale to assess identification risk

For the purposes of this article, it is considered that to comply with the commitment to preserve statistical confidentiality, it is useful to develop a scale to assess the risk of identification in the information prior to its dissemination. The scale must allow at least:

1. Meet the requirements indicated in section 2 of this article.
2. Characterize different levels of identification risk in the public information.
3. That the different risk levels specify the characteristics of the information as well as the circumstances that correspond to it.
4. That its application allows statistical agencies to find a balance between their role of providing quality information and preserving statistical confidentiality.

Additionally, it is necessary to complement the rating scale with some operating policies that indicate:

a) The moment in which it must be applied.
b) The decisions to be taken according to each level of risk.

In consideration of the previous premises, the following scale is proposed:

Identification risk level	Description
High	Identification is immediate. Just by accessing the information it is possible to recognize the person or organization to which the data corresponds; or when the identification is deduced from the combination of different data represented in the same product through which the information is presented.
Medium	Identification is achieved by adding or subtracting some classes or groupings from the same tabulation and combining the result with other statistical or geographical information products.
Low	Identification is achieved by combining the information with different public and private data repositories using analytical techniques, software, and computer equipment.
Null	It is not possible to carry out the identification by any means or by the combination of any information.

Operation policies:

a) The scale can be applied to statistical information as well as georeferenced statistical information.
b) The scale should be used when the official statistical information is analyzed before it is disseminated; whoever performs the analysis must associate the information with the level of risk that best describes its attributes and conditions.
c) When the risk of identification is high or medium, it is recommended not to disseminate the information with the current attributes and conditions, and to identify alternatives to publicize the results of the processing in a different way that does not compromise the anonymity of data providers.
d) The scale must be applied each time the information is modified prior to dissemination.

The application of the scale proposal does not exclude automated analysis through any computer tool. Finally, it is recommended that the level of risk identified and the arguments for each decision have to be documented and to form part of the rest of information´s production documentation to generate and preserve knowledge in statistical agencies.

4 Conclusions

The scale proposal represents an initiative to formalize and standardize in statistical agencies the way in which the information is reviewed prior to its dissemination. The levels and descriptions that make up the scale were written in a generic way; each national statistical agency can modify them according to the context and legal requirements that apply. It should not be forgotten that the scale is not an end, but a means to fulfill the mission of providing quality information and preserving statistical confidentiality.

5 References

Hernández González, Sergio (2005) *Historia de la estadística*. En Revista de Divulgación Científica y Tecnológica de la Universidad Veracruzana, Volumen XVIII, Número 2. Available in: https://www.uv.mx/cienciahombre/revistae/vol18num2/articulos/historia/

Masciadri, Viviana (2013). *Nombre y apellido o razón social, domicilio y rama de actividad: ¿deben o no exceptuarse del secreto estadístico? Una revisión comparativa. Espacios Públicos*, 16(37),141-174.[fecha de Consulta 29 de Septiembre de 2021]. ISSN: 1665-8140. Available in: https://www.redalyc.org/articulo.oa?id=67628073008

ONU (2014) *UN Fundamental Principles of Official Statistics*. Available in: https://unstats.un.org/unsd/dnss/hb/S-fundamental%20principles_A4-WEB.pdf

United Nations National Quality Assurance Frameworks Manual for Official Statistics (UN NQAF Manual). Available in: https://unstats.un.org/unsd/methodology/dataquality/un-nqaf-manual/

UN Fundamental Principles of Official Statistics – Implementation guidelines. Available in: https://unstats.un.org/unsd/dnss/gp/Implementation_Guidelines_FINAL_without_edit.pdf

Quintela del Río, Eduardo (2019, septiembre, 03) *Estadística Básica Edulcorada*. Available in: https://bookdown.org/aquintela/EBE/introduccion.html

Rodríguez Castilla, Amadeo (2007) *Experiencias Importantes en la Historia de los Censos, y el Censo General 2005 de Colombia*. En Revista de la Información Básica. Vol 1. No. 2. Available in: https://sitios.dane.gov.co/revista_ib/html_r4/articulo5_r4.htm