Anonymization of Anual Agricultural survey (AAS) 2020

UBOS

2022-11-30

Introduction

[Describe the survey by providing the main objectives, the number of round and/or seasons, the sampling design and the period of data collection for the different rounds and season. These information can be taken from the report of the survey if already published.]

- Datasets of the survey

[Remind the commitment of the institute on disseminating microdata and recall the need of the need of SDC to protect respondent's confidentiality]

Disclosure scenario analysis

Anonymization or Statistical Disclosure control is a process of treating data to reduce the risk of disclosure. In the AHS, the main unit that need to be protected regarding ethics and legislation in the Agriculral holding (family holding and entreprises). However, these holding can be indirectly disclosed throughout some linked units such as Parcels, Crops, Livestocks (when we disclose a percel in the microdata, we can know the holding it belongs to and thus disclose information of that holding). In this regards, beyond preventing the disclosure of holdings, we need to prevent the disclosure of parcels, crops and livestocks.

The units in the microdata can be disclosed in two ways:

- Identity disclosure: Identity disclosure can happen if an intruder succeeds in matching records in the released microdata with external datasets containing identifying information (region, age, sex, name, or any other variable in the released microdata that may exist in an external register).
- Attribute disclosure: intruder can identify a record without doing a linkage. This is known as attribute disclosure. One special situation is the Nosy neighbor scenarios. These scenarios assume the intruder has enough information on a unit or units; this information stems from his/her personal knowledge. In other words, the intruder belongs to the circle of acquaintances of a statistical unit.

The objective of the disclosure scenario is to identifying variables and sensitives variables that the intruder can used to disclose information. This, particularly, depends on the environment where the microdata is released.

The disclosure scenario exercice have been done with colleagues from Geostat and all the variables in the differents datasets have been classified in on of the below categories:

- D: variables to delete (not useful for the user)
- Q: quasi-indentifier (to be anonymized)
- direct identifier: to be deleted (name, adresse, Phone, etc.)
- S: Sensitive variables
- L: Variable that are linked to a quasi-identifier
- ID variable: identification code
- -: Variable that does not need SDC and should be disseminated as they are

Data processing before anonymization

[Describe the main task of which the data pre-processing consisted]

Disclosure risk assessment methods

Assessing the disclosure risk is one of the key steps in SDC process. For assessing the Disclosure risk of the 2020 AHS, we resort to the follwing methods:

categorical quasi-identifers

- **k-anonymity**

to highlight uniqueness of records regarding the combination of categorical quasi-identifiers

- **Special Uniques Detection Algorithm (SUDA)**

to go beyond uniqueness of the combination of quasi-identifiers and look at the uniqueness in the subset of combination of quasi-identifer (Special Unique)

- **Probabilistic risk**

Assess the disclosure risk as a probability. Units have the same combination of keys will have the same risk * extected number of re-identification: a global risk measure

for continous quasi-identifers

In the literature, the risk of continous variable are calculated a posteriori. For

Given the nature of the data structure, we will lay emphasis on the outlierliness of observation regarding the continous variables.

- Gini index: to assess the magnitude of inequality in some variables (Production, Area, Save vlaues, etc.)
- Lorenz-curveThe Lorenz curve displays the deviations of the empirical distribution from a perfectly equal distribution as the difference between two graphs
- A customized metric we developped to detect the right-skewedness of continous variables:skewness function was described by Groeneveld, R. A. and Meeden ($right_skewedness = \lambda(\alpha) = \frac{Q(\alpha)+Q(1-\alpha)-2\times median}{Q(1-\alpha)-Q(\alpha)}$) computed for values that are above the statistical mode (M)

$$RS = \lambda_{\{x_i>M\}}(\alpha) = \frac{Q_{\{x_i>M\}}(\alpha) + Q_{\{x_i>M\}}(1-\alpha) - 2 \times median_{\{x_i>M\}}}{Q_{\{x_i>M\}}(1-\alpha) - Q_{\{x_i>M\}}(\alpha)}$$

with $\alpha = 0.05$ This skewedness coefficient varies from -1 tp +1. Values close to +1 indicated the the presence of a very skewed right tail in the distribution of the variable.

Other anonymization measures of households