

» Synthetic data as Public Use Files:

An application to the Household Budget Survey «

Inês Rodrigues (ines.rodrigues@ine.pt)

Methodology Unit | Statistics Portugal

 11 July 2019

»



Summary



- What are Public Use Files (PUF)?
- Producing PUF by generating synthetic data
- Producing PUF for the Household Budget Survey
- Results and discussion



Public Use Files (PUF)



REGULATION (EC) No 223/2009 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL
of 11 March 2009

Article 19

Public use files

Persons, households,
enterprises, local units

Statistical Disclosure
Control applied to some
degree

Data on individual statistical units may be disseminated in the form of a public use file consisting of anonymised records which have been prepared in such a way that the statistical unit cannot be identified, either directly or indirectly, when account is taken of all relevant means that might reasonably be used by a third party.

Through direct
identifiers (e.g. names,
ID numbers)

Through indirect
identifiers (e.g. sex,
age, region)



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

3/17 »

Synthetic data as PUF



- Methods for protecting confidentiality in microdata:

Masking methods

- » Modify the original data
- » Aggregate or suppress data (non-perturbative methods)
- » Apply an element of error to data (perturbative methods)

Synthetic data generation

- » Replace the original data
- » Most relevant statistical properties are kept



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

4/17 »



Synthetic data as PUF



- D : original microdata set
- Y : vector of k variables from D , whose relationships are intended to be preserved

Parametric approach:

Joint conditional density of Y_1, Y_2, \dots, Y_k – Raghunathan et al., 2001:

$$f(Y_1, \dots, Y_k | X, \theta_1, \dots, \theta_k) = f_1(Y_1 | X, \theta_1) \prod_{v=2}^k f_v(Y_v | X, Y_1, \dots, Y_{v-1}, \theta_v)$$

- » Model each conditional distribution a given appropriate regression model (e.g., linear, logistic or log-linear regression)



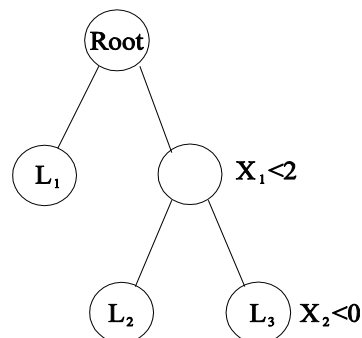
Synthetic data as PUF



Non-parametric approach:

CART (classification and regression trees) algorithm – Reiter (2005), Nowok et al., (2017):

- » Recursive partition of the original dataset (based on *yes/no* questions on the predictors)
- » Groups have increasingly homogeneous outcome, Y_j
- » In each final group (leaf), values approximate the conditional distribution of Y_j
- » Synthetic values are generated by sampling from the appropriate leaf
- » CART can be used to simulate each variable sequentially, by conditioning on already generated variables, as in the parametric approach



Example of a tree structure (source: Reiter, 2005)



Disclosure risk



Loong et al. (2013):

ID	Indirect identifiers				Confidential variable(s)	
	Sex	Age	Region	...	Income	...
1	M	54	3	...	26 585	...
2	M	22	4	...	3 345	...
3	F	49	3	...	13 456	...
...

- We assume the user knows which units are included in D and their values regarding m indirect identifiers.
- Based on this m variables, the user attempts to obtain information on a confidential variable, T , from the synthetic microdata set D' .



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

7/17 »



Disclosure risk



Expected match risk (EMR):

$$\text{EMR} = \sum_{i=1}^{n_R} \frac{Z_i}{C_i}$$

w_{iq} : value of the indirect identifier q for unit i in D ($i = 1, \dots, n_R$; $q = 1, \dots, m$)

w_{jq} : value of the same identifier for unit j in D' ($j = 1, \dots, n_S$; $q = 1, \dots, m$)

» $R_{ij} = 1$ if $w_{iq} = w_{jq} \forall q$ ($q = 1, \dots, m$); $R_i = 0$ otherwise

» $C_i = \sum_{j=1}^{n_S} R_{ij}$

» $U_{ic} = 1$ if $t_i = t_c$; $U_{ic} = 0$ otherwise (for categorical T)

» $Z_i = \sum_{c=1}^{C_i} U_{ci} \rightarrow$ Total number of records that are a real match for unit i in D



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

8/17 »



Disclosure risk



Expected match risk (EMR):

$$\text{EMR} = \sum_{i=1}^{n_R} \frac{Z_i}{C_i}$$

True match risk (TMR):

$$\text{TMR} = \sum_{i=1}^{n_R} K_i$$

- » $I_i = 1$ if $Z_i > 0$; $I_i = 0$ otherwise
- » $K_i = 1$ if $C_i = 1 \wedge I_i = 1$; $K_i = 0$ otherwise

- EMR reflects the chance of an user randomly establishing a true match for each unit i in D
- TMR reflects the chance of an user correctly and uniquely identifying each unit i in D



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

9/17 »

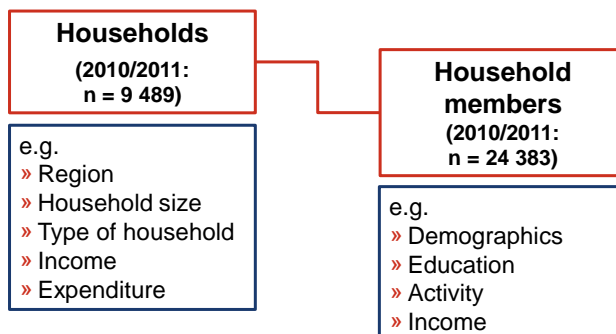


PUF for the Household Budget Survey



- Household Budget Survey (HBS) aims at producing data on consumption expenditure

Microdata:



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

10/17 »



PUF for the Household Budget Survey



Step 1 – Background variables:

- Region (NUTS II) and household size are generated as multinomially distributed random numbers
- Synthetic data has the same number of households as real data (SUF)

Step 2 – Sequential regressions:

- Both the parametric and non-parametric approaches are applied, where:
 - » X = Region (NUTS II) and household size
 - » Y = Type of dwelling, income and expenditure totals and the main identifying variables at the individual level (country of birth, country of citizenship, marital status, level of studies completed, status in employment and economic sector in employment)



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

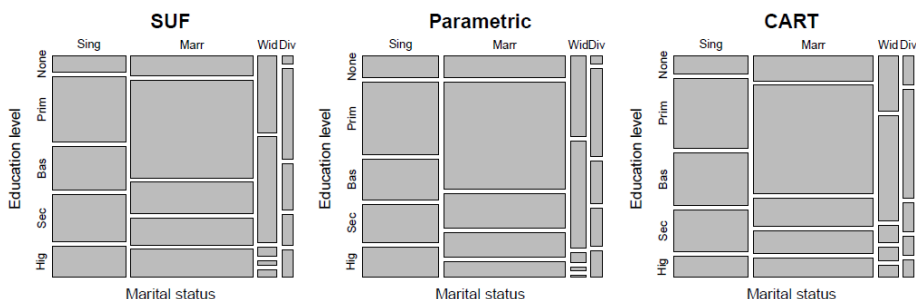
11/17 »



Results and discussion



- » Weighted frequency distribution of education level by marital status, in the real (SUF) and synthetic HBS datasets

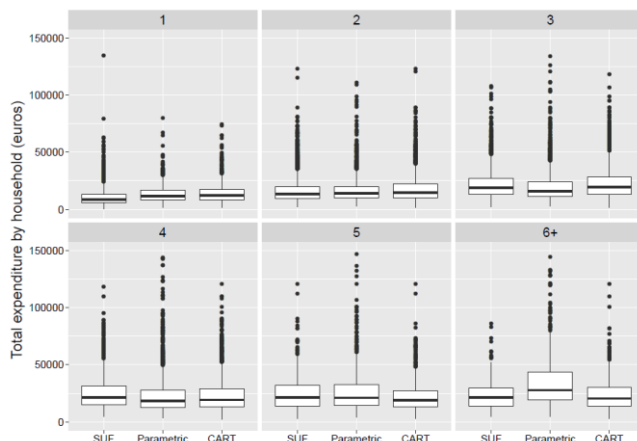


INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

12/17 »

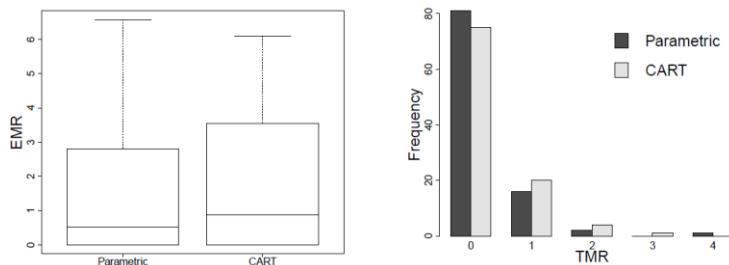
Results and discussion

» Total annual consumption expenditure by household size, in the real (SUF) and synthetic HBS datasets



Results and discussion

» EMR and TMR distributions ($r = 100$, $m = 6$ (sex, age, HH size, marital status, status in employment and country of citizenship), T are income and expenditure totals and $p = 0.05$).





Results and discussion



- High utility from both approaches:

- » Relevant relationships between variables are kept;
- » Results obtained regarding the main statistics computed from HBS data:

Data	Median equivalised disposable income (€)	At-risk-of-poverty rate (after social transfers) (%)	Gini coefficient of equivalised disposable income (%)	Annual mean consumption expenditures (€) of households
SUF	11 000	14.8	33.2	20 391
Param	11 140	19.2	31.7	19 942
CART	10 800	15.5	32.6	19 661

- However, the additional flexibility from CART results in a slight increase in disclosure risk.



References



- Drechsler, J. and Reiter, J.P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, 55, 3232 – 3243.
- Franconi, L. *International access through Public Use Files*. OECD Expert Group for International Collaboration on Microdata Access – Final Report. Paris, July 2014.
- Loong, B., Zaslavsky, A.M., He, Y. and Harrington, D.P. (2013). Disclosure control using partially synthetic data for large-scale health surveys, with applications to CanCORS. *Statistics in Medicine*, 32, 4139 – 4161.
- Nowok, B., Raab, G.M. and Dibben, C. (2017). Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R. *Statistical Journal of the IAOS*, 33, 785 – 796.
- Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J. and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology*, 27, 85 – 95.
- Reiter, J.P. (2005). Using CART to Generate Partially Synthetic, Public Use Microdata. *Journal of Official Statistics*, 21, 441 – 462.



Thank you!

Inês Rodrigues (ines.rodrigues@ine.pt)