Disclosure Risk of Spatial Data

Definition and recommendations

Maëlle Fontaine

INSEE (France)

Disclosure Risk of Spatial Data

1. Contextual elements

2. Disclosure risk: what is it and how to deal with it?

3. Measuring disclosure risk while taking spatial features into account

4. Recommendations

Contextual elements

Contextual elements

More and more statistical sources are geocoded

INSs are encouraged to disseminate grid data because:

- it gets rid of administrative zoning
- it allows comparisons over countries and over time

A trade-off to find between risk and utility must be found

- spatial data involves more risk of re-identification
  -> need to distort data
- but it offers many possibilities of analysis (high utility)
  -> not distort data too much

Contextual elements

Various initiatives have already been undertaken

- several tests within the framework of Eurostat Grant *"Harmonized protection of 2021 Census data"*
- extensions of these tests in a chapter dedicated to confidentiality of spatial data, within the Eurostat *"Handbook of spatial analysis"*

More details in the forthcoming publications!

Disclosure risk: what is it and how to deal with it?

Disclosure risk: what is it?

General considerations about disclosure risk

- disclosure occurs when an intruder uses released data to learn some information he does not already know
- disclosure is regulated by law or conventions (thresholds)
- there is no universal measure of disclosure risk
- differencing issues (interactions with other zonings)

Disclosure risk: how to deal with it?

How to prevent disclosure?

- considering different kinds of users (general public VS experts)
- aggregating the data to a superior geographic level
- perturbating the data, *i.e.* applying a Statistical Disclosure Control (SDC) method

Traditionnally

- SDC methods are aspatial
- spatial correlations can be very distorted
- differencing issues are treated separately

Disclosure risk: how to deal with it?

2 families of SDC methods

	Pre-tabular methods	Post-tabular methods
Data perturbated	micro-data	tabular data
Example	swapping	rounding
Pros	less distortion of correlations	easier to implement
Cons	impression that nothing is done	hard to keep consistency

SDC methods are under constraints:

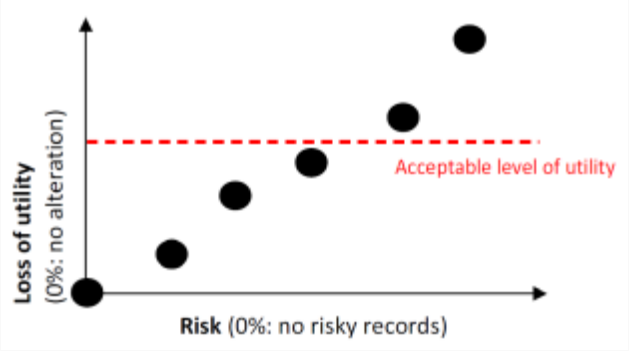
- keeping empty cells empty
- keeping margins
- consistency from one table to another …

Measuring disclosure risk while taking spatial features into account

How to take spatial features into account?

Recommanded approach

1. at the record level, calculating an individual score which depends on the local context
2. setting a threshold as a quantile of the score distribution
3. applying a SDC method targeting the riskiest records (above the threshold) and taking geography into account, *e.g.* Targeted Record Swapping (TRS)
4. quantifying the loss of utility through a set of distortion indicators (average absolute deviation, Moran's I…)
5. repeating the operation for several thresholds
6. drawing a risk-utility map to choose the best trade-off



Example: measuring disclosure risk at record level

Shlomo Tudor Groom, 2010 (ONS Census)

- use 3 nested geographic levels
- compute a score for each individual and each geographic level, based on frequency counts of univariate distributions

Frequency counts of a set of key variables for a given geographic level l

$g(region)$	Women	Men	<25	25-49	>=50
1	5	6	1	7	3
2	4	7	3	9	2
...					
G	11	0	2	6	3

woman / 30 y.o. / g=1: score=$(1/5 + 1/7)/2 \simeq 0.17 < T_l$ -> not risky for level l

Nagy, 2015 (Hungary Census)

- applies a pre-tabular method to grid data
- introduces multivariate distributions

Example: extensions of tests carried out for the Eurostat Grant

Buron Fontaine, 2017 (France, experimental tests about grid data)

- use a *ad hoc* hierarchical partition of a French region into 3 nested levels
    1. small rectangles (at least 100 individuals)
    2. big rectangles (at least 5000 individuals)
    3. NUTS3
- compute individual scores as in example 1
- apply targeted record swapping between similar records, belonging to the same superior hierarchical level
- make a risk-utility analysis
- main result: no distortion of spatial correlations for variables taken into account in the swapping, or correlated to them

Recommendations

Recommendations

Recommendations if variables to release are sensitive:

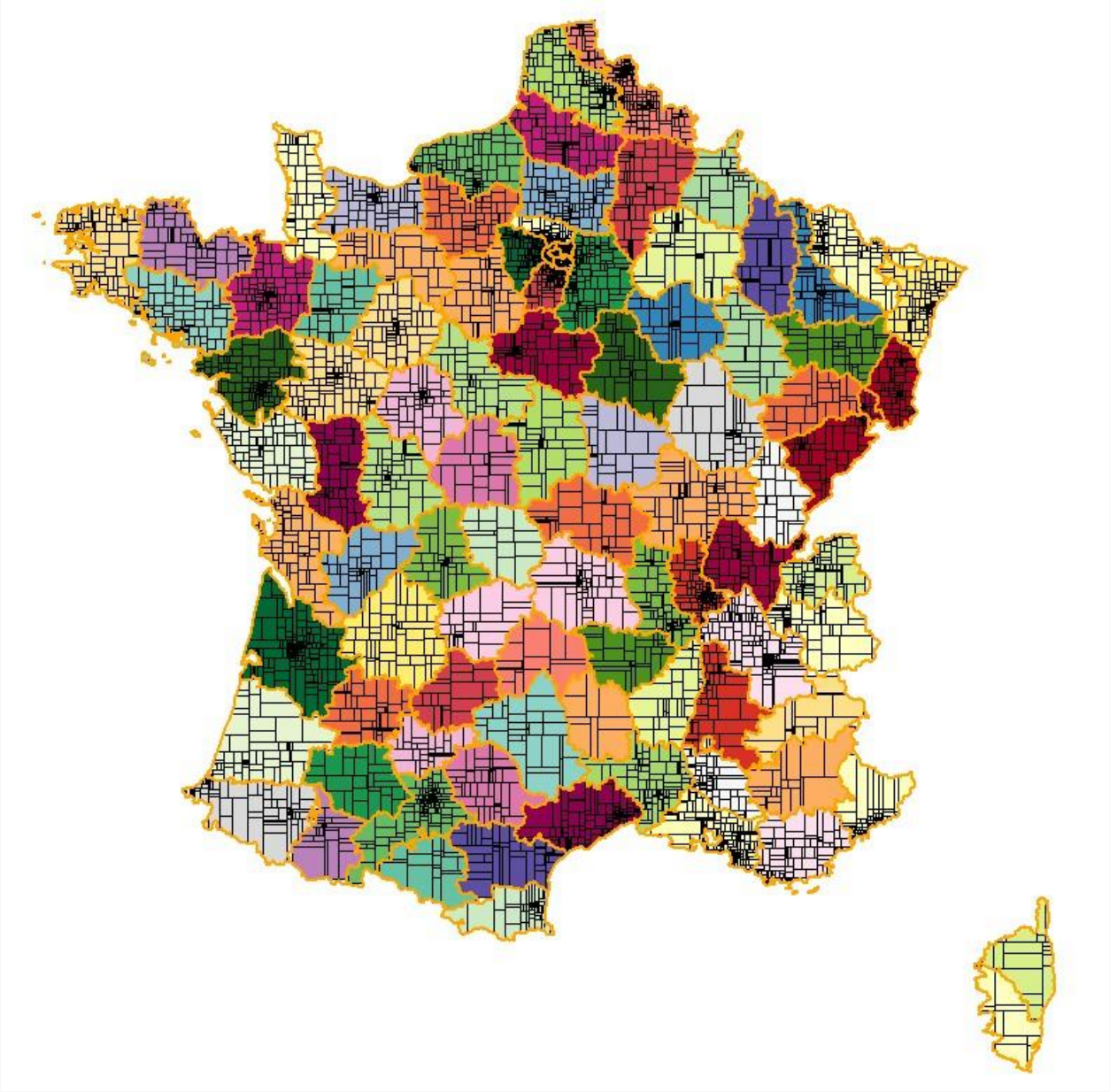
- considering pre-tabular methods like swapping because it scrambles the data while conserving spatial correlations …
- … but combining it with another method to guarantee mechanically thresholds rules and keep a reasonable level of perturbation
- keeping algorithms flexible
- anticipating computing times issues (eventually considering simplified versions)
- systematizing risk-utility analysis
- dealing separately with differencing problems

Conclusion

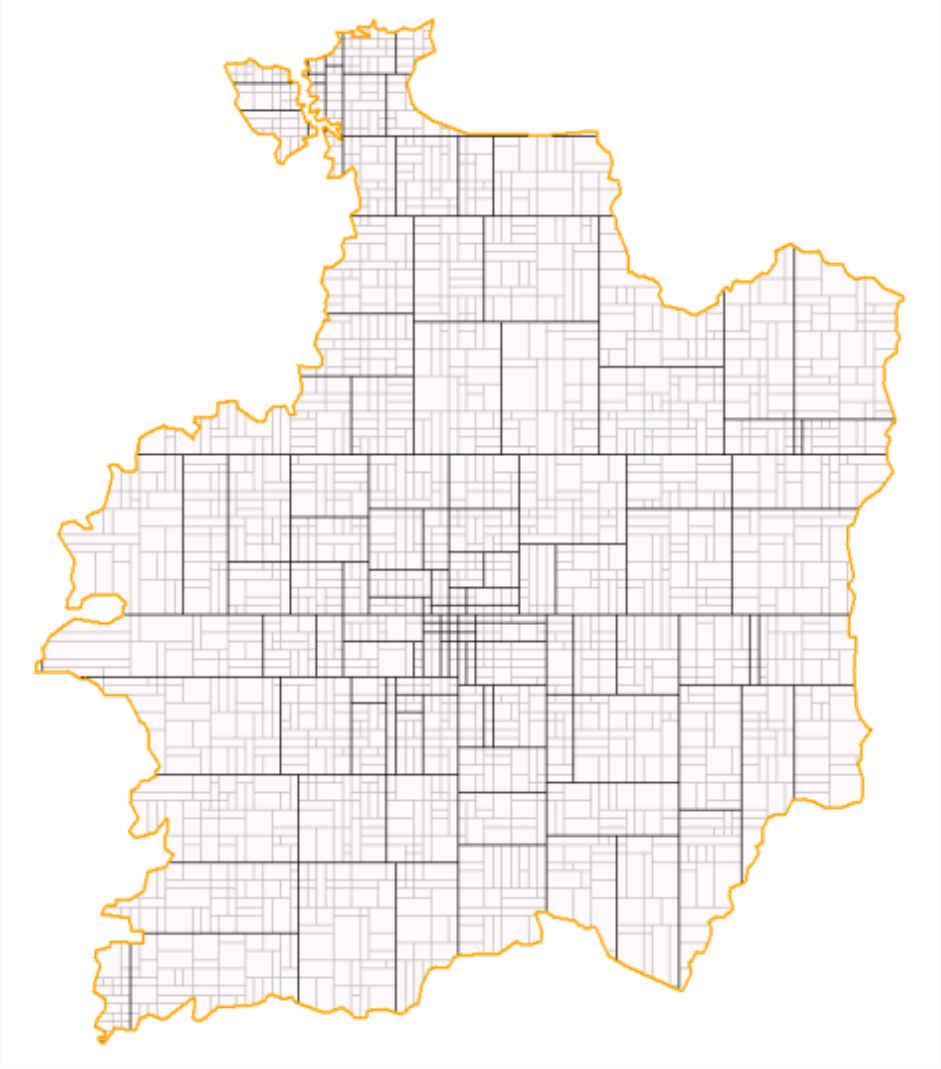
Thank you for your attention!

Some questions ?

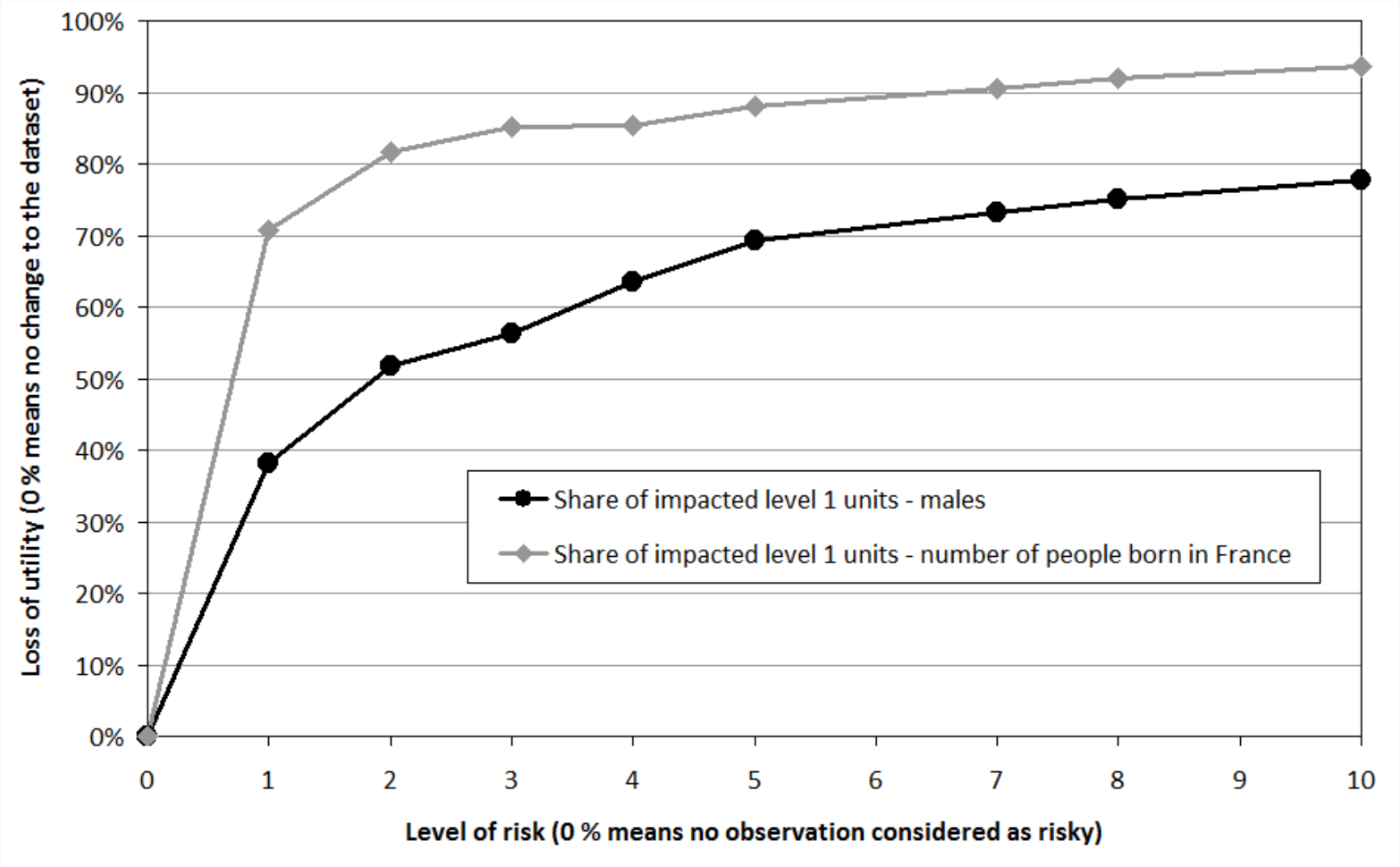
Appendix



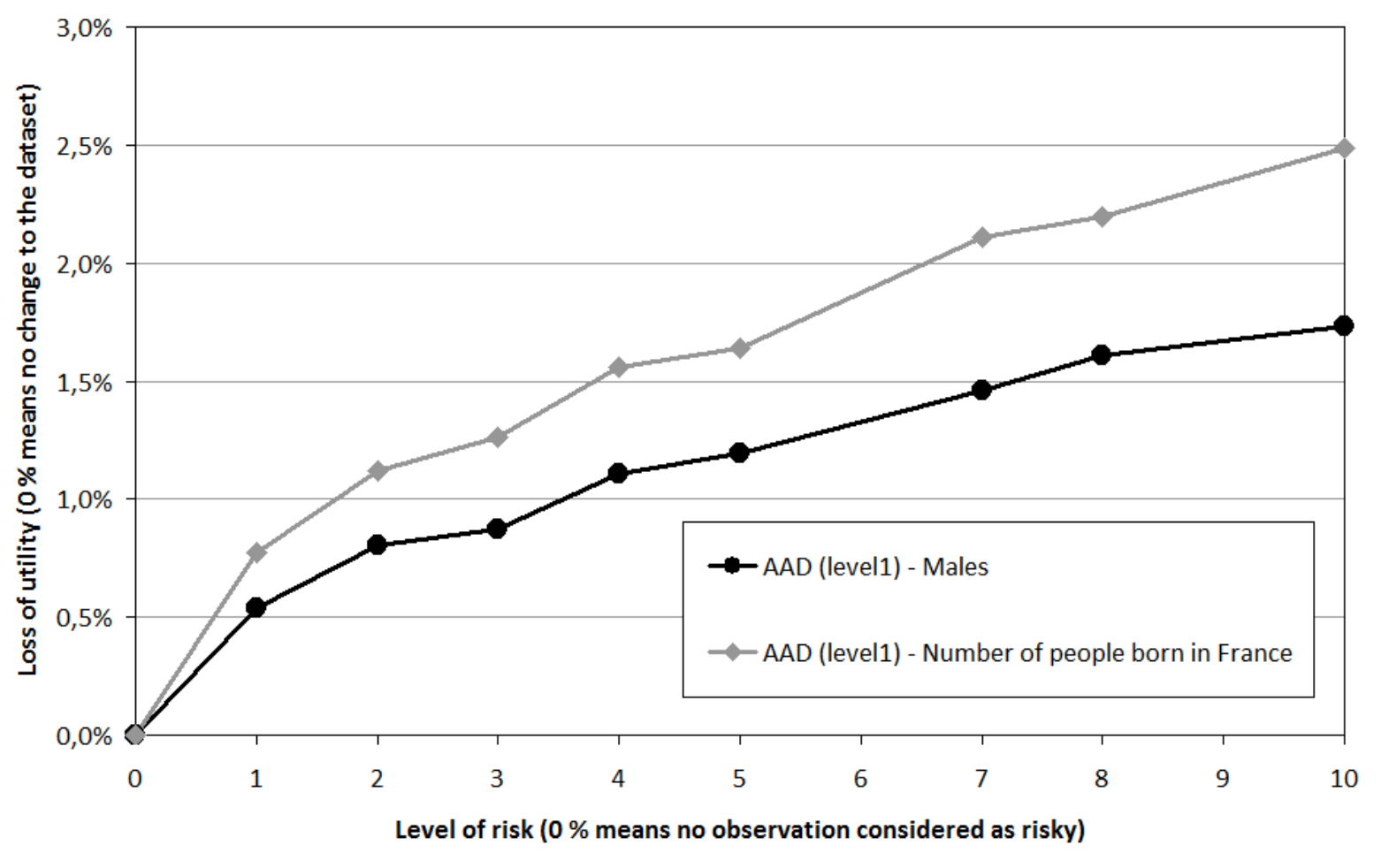
Appendix



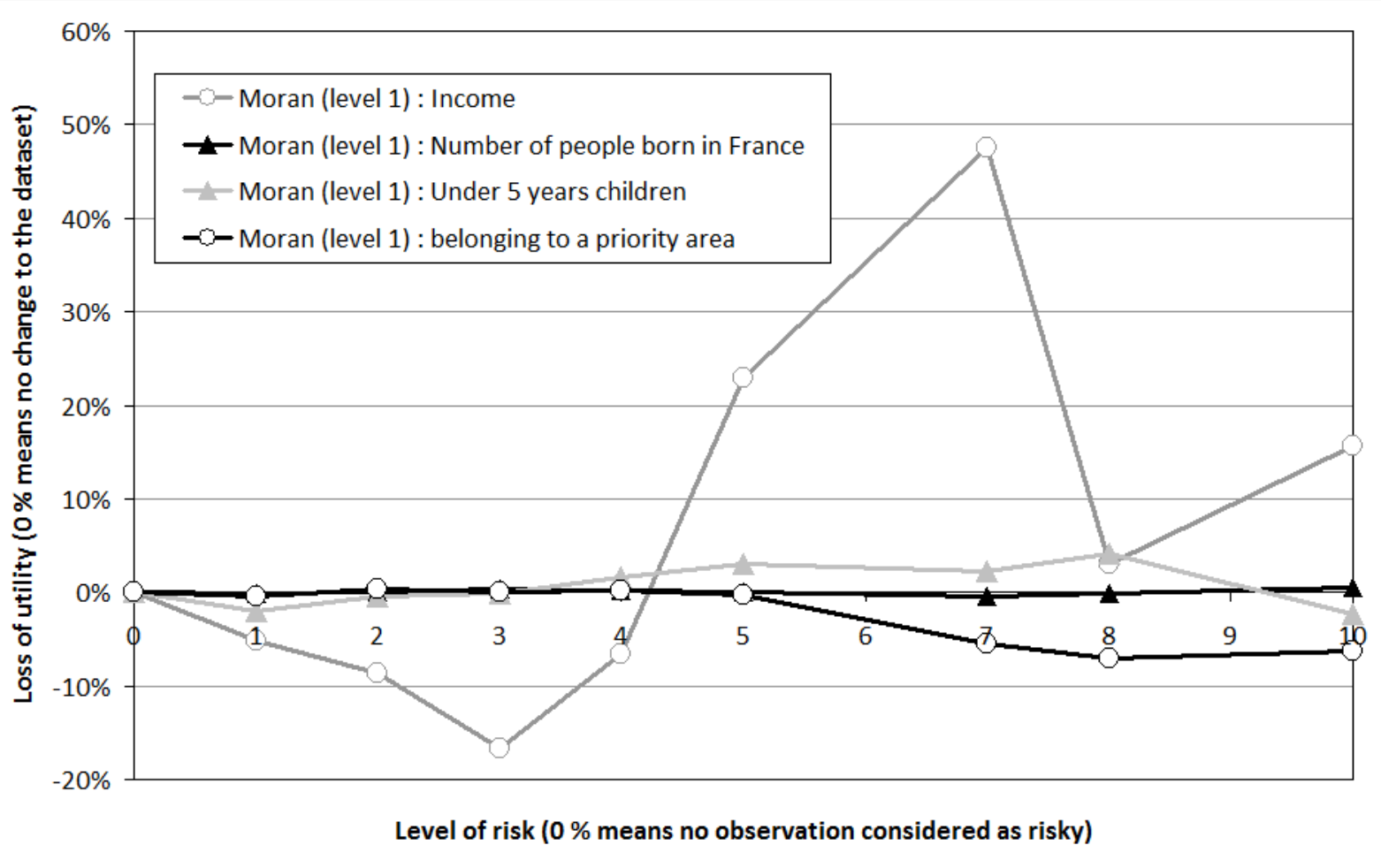
Appendix


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Loss of utility (0 % means no change to the dataset)
0
1
2
3
4
5
6
7
8
9
10
Level of risk (0 % means no observation considered as risky)
Share of impacted level 1 units - males
Share of impacted level 1 units - number of people born in France

Appendix


3,0%
2,5%
2,0%
1,5%
1,0%
0,5%
0,0%
Loss of utility (0 % means no change to the dataset)
0
1
2
3
4
5
6
7
8
9
10
Level of risk (0 % means no observation considered as risky)
AAD (level1) - Males
AAD (level1) - Number of people born in France

Appendix


Moran (level 1) : Income
Moran (level 1) : Number of people born in France
Moran (level 1) : Under 5 years children
Moran (level 1) : belonging to a priority area
60%
50%
40%
30%
20%
10%
0%
-10%
-20%
0
1
2
3
4
5
6
7
8
9
10
Loss of utility (0 % means no change to the dataset)
Level of risk (0 % means no observation considered as risky)