3rd IAB Workshop on Confidentiality and Disclosure

On Constrained Microaggregation

Vicenç Torra[1]

20-21 November 2008

[1] Institut d'Investigació en Intel·ligència Artificial (IIIA-CSIC)

Introduction

- Edit constraints
- ... and microaggregation

Introduction

- When data is edited, variables satisfy some constraints,
- Application of masking methods,
  ... causes the violation of the constraints

Introduction

- Is microaggregation appropriate ?
- Constrained microaggregation.

Outline

- Edit constraints
- Microaggregation
- Microaggregation and Edit Constraints
  - Linear Constraints
  - Nonlinear Constraints
  - Constraints on the Values
  - One variable governs another
  - Restriction on the values
- Conclusions

Edit Constraints

Introduction

- Edit constraints
  - A classification of the constraints

Edit Constraints

- Constraints on the possible values.
  - Values restricted to a predefined set
    - Values in a interval:

$$\text{EC-PV: } age \in [0, 125]$$

  - Generalizable for subsets of variables
    - Values (v_1, v_2) in a subset of $D_1 \times D_2$

Edit Constraints

- One variable governs the possible values of another one
  - The values of a variable v_2 constrained by v_1
    - E.g., variable *sex* governing *number of pregnacies*

      EC-GV1: If *sex=male* THEN *number of pregnacies = 0*
    - or, e.g.[1]:

      EC-GV2: IF *age* < 17 THEN *gross income* $<$ *mean income*
    - or, e.g.[2]

      EC-GV3: *harvested acres* $\leq$ *planted acres*

[1] Shlomo, N., De Waal, T. (2008), Protection of micro-data subject to edit constraints against statistical disclosure, Journal of Official Statistics 24:2 229-253.

[2] Pierzchala, M. (1994) A review of the state of the art in automated data editing and imputation, in Statistical Data Editing, Vol. 1, Conference of European Statisticians Statistical Standards and Studies N. 44, United Nations Statistical Commission and Economic Commission for Europe, 10-40.

Edit Constraints

- Linear constraints.
  - Some variables satisfy some linear relationships.
    - E.g., *gross* in terms of *net* and *tax*

$$\text{EC-LC1: } \mathit{net} + \mathit{tax} = \mathit{gross}$$

Edit Constraints

- Non-linear constraints.
  - The relationship between variables is not linear.
    * Relationship between *applicable VAT Rate*, *price exc. VAT*, and *retail price*:

      EC-NLC1: *price exc. VAT* $\cdot$ (*1.00* + *applicable VAT Rate*) = *retail price*
    * Relationship between *wage sum*, *hours paid for*, and *wage rate*[3]:

      EC-NLC2: *wage sum* = *hours paid for* $\cdot$ *wage rate*

[3] Gasemyr, S. (2005) Editing and imputation for the creation of a linked micro file from base registers and other administrative data, Conference of European Statisticians, WP8.

Edit Constraints

- Other types of constraints.
  - E.g. constraints on categorical (ordinal or nominal) variables

Edit Constraints

- Values are restricted to exist in the domain
  - Values not only in the range but also exist in the data.
    - E.g. ages really existing in the population
      $\longrightarrow$ not enough to be in [0,125].
  - A perturbative method applied to data with ages in [0,30] should not lead to a file with a value equal to 50.
    - Application in linked files.

Microaggregation

Microaggregation

- Microaggregation: a perturbative method
  - Notation.
    - $u_{ij} \in \{0, 1\}$ a partition: $u_{ij} = 1$ iff record j is assigned to the ith cluster.
    - v_i represents the ith cluster
    - k minimum number of records in a cluster, g number of clusters.
  - Formalization.

$$
\begin{array}{ll}
\text{Minimize} & SSE = \sum_{i=1}^{g} \sum_{j=1}^{n} u_{ij} (d(x_j, v_i))^2 \\
\text{Subject to} & \sum_{i=1}^{g} u_{ij} = 1 \text{ for all } j = 1, \ldots, n \\
 & 2k \geq \sum_{j=1}^{n} u_{ij} \geq k \text{ for all } i = 1, \ldots, g \\
 & u_{ij} \in \{0, 1\}
\end{array}
$$

Microaggregation

- Microaggregation. The Operational approach.
  - 1. Clustering:
    Partition the set of records
    $\longrightarrow$ each partition element should have at least k records
  - 2. Cluster representatives:
    Compute a cluster representative for each cluster
  - 3. Replacement:
    Replace each record by its cluster representative

Microaggregation

- Microaggregation and k
  - The larger the k, the smaller the risk.
- Microaggregation and k-anonymity
  - k-anonymity: k-indistinguishable records
  - Satisfied when all variables microaggregated together
    $\rightarrow$ microaggregation on the $\mathbb{R}^m$ space
  - Otherwise, in general, not satisfied.

Microaggregation and Edit Constraints

Linear Constraints

Microaggregation and the edit constraints

- Microaggregation can deal easily with edit constraints
- Notation:
  - $x_1, \ldots, x_n$ records
  - $V_1, \ldots, V_m$ variables
  - $x_{i,j}$: value of record x_i for variable V_j

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - Simplification on notation: V in terms of $V_1, \ldots, V_K$

V	V_1	$\ldots$	V_K
x_1	$x_{1,1}$	$\ldots$	$x_{1,K}$
$\vdots$	$\vdots$		$\vdots$
x_N	$x_{N,1}$	$\ldots$	$x_{N,K}$

  - Assumption$_1$: All the variables in the linear model are microaggregated together.
  - Assumption$_2$: Steps 1, 2, and 3 of the operational approach can be separated.

    $\rightarrow$ cluster representative for each cluster satisfying the constraint

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - Simplification on notation: V in terms of $V_1, \ldots, V_K$

V	V_1	$\ldots$	V_K
x_1	$x_{1,1}$	$\ldots$	$x_{1,K}$
$\vdots$	$\vdots$		$\vdots$
x_N	$x_{N,1}$	$\ldots$	$x_{N,K}$

  - Assumption$_3$: Linear constraint of the form $V = \sum_{i=1}^{K} \alpha_i V_i$
  - Naturally, the data also satisfies the constraints (i.e., the data were already edited). I.e.,
    $x_j = \sum_{i=1}^{K} \alpha_i x_{j,i}$ for all j.

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - Simplification on notation: V in terms of $V_1, \ldots, V_K$

V	V_1	$\cdots$	V_K
x_1	$x_{1,1}$	$\cdots$	$x_{1,K}$
$\vdots$	$\vdots$		$\vdots$
x_N	$x_{N,1}$	$\cdots$	$x_{N,K}$
$\mathbb{C}(x_1, \ldots, x_N)$	$\mathbb{C}(x_{1,1}, \ldots, x_{N,1})$	$\cdots$	$\mathbb{C}(x_{1,K}, \ldots, x_{N,K})$

  - Assumption$_4$: The cluster representative is a function of the data in the cluster (each variable, independently): $\mathbb{C}$

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - Simplification on notation: V in terms of $V_1, \ldots, V_K$

V	V_1	$\cdots$	V_K
x_1	$x_{1,1}$	$\cdots$	$x_{1,K}$
$\vdots$	$\vdots$		$\vdots$
x_N	$x_{N,1}$	$\cdots$	$x_{N,K}$
$\mathbb{C}(x_1, \ldots, x_N)$	$\mathbb{C}(x_{1,1}, \ldots, x_{N,1})$	$\cdots$	$\mathbb{C}(x_{1,K}, \ldots, x_{N,K})$

  - From these assumptions, we require:

$$\mathbb{C}(x_1, \ldots, x_N) = \sum_{i=1}^{K} \alpha_i \mathbb{C}(x_{1,i}, \ldots, x_{N,i})$$

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - Simplification on notation: V in terms of $V_1, \ldots, V_K$

V	V_1	$\cdots$	V_K
x_1	$x_{1,1}$	$\cdots$	$x_{1,K}$
$\vdots$	$\vdots$		$\vdots$
x_N	$x_{N,1}$	$\cdots$	$x_{N,K}$
$\mathbb{C}(x_1, \ldots, x_N)$	$\mathbb{C}(x_{1,1}, \ldots, x_{N,1})$	$\cdots$	$\mathbb{C}(x_{1,K}, \ldots, x_{N,K})$

  - As $x_j = \sum_{i=1}^{N} \alpha_i x_{j,i}$ for all j in $\{1, \ldots, N\}$, we write:

$$\mathbb{C}(\sum_{i=1}^{K} \alpha_i x_{1,i}, \ldots, \sum_{i=1}^{K} \alpha_i x_{N,i}) = \sum_{i=1}^{K} \alpha_i \mathbb{C}(x_{1,i}, \ldots, x_{N,i})$$

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - Simplification on notation: V in terms of $V_1, \ldots, V_K$

V	V_1	$\cdots$	V_K
x_1	$x_{1,1}$	$\cdots$	$x_{1,K}$
$\vdots$	$\vdots$		$\vdots$
x_N	$x_{N,1}$	$\cdots$	$x_{N,K}$
$\mathbb{C}(x_1, \ldots, x_N)$	$\mathbb{C}(x_{1,1}, \ldots, x_{N,1})$	$\cdots$	$\mathbb{C}(x_{1,K}, \ldots, x_{N,K})$

  - We also require reflexivity:

$$\mathbb{C}(x, \ldots, x) = x$$

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - Proposition 1. (proof based on Functional Equations[4])

    $\mathbb{C}$ a function satisfying

    $$\mathbb{C}(\sum_{i=1}^{K} \alpha_i x_{1,i}, \ldots, \sum_{i=1}^{K} \alpha_i x_{N,i}) = \sum_{i=1}^{K} \alpha_i \mathbb{C}(x_{1,i}, \ldots, x_{N,i})$$

    for given values $\alpha_1, \ldots, \alpha_K$ ($\alpha_i \neq 0$) and arbitrary values $x_{i,j}$ for $1 \leq i \leq N$ and $1 \leq j \leq K$, and reflexivity

    $$\mathbb{C}(x, \ldots, x) = x$$

    Then, the most general solution for $\mathbb{C}$ is a function of the form

    $$\mathbb{C}(x_1, \ldots, x_N) = \sum_{i=1}^{N} \kappa_i x_i$$

    for κ_i such that $\sum_{i=1}^{N} \kappa_i = 1$ but otherwise arbitrary.

[4] Aczél, J. (1987) A Short Course on Functional Equations; J. Aczél (1966) Lectures on Functional Equations and their Applications, Academic Press.

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - Proposition 2.
    $\mathbb{C}$ as before, but valid for all $\alpha_1, \ldots, \alpha_K$ ($\alpha_i \neq 0$):
    Same result:
    Then, the most general solution for $\mathbb{C}$ is a function of the form
    $$\mathbb{C}(x_1, \ldots, x_N) = \sum_{i=1}^{N} \kappa_i x_i$$
    for κ_i such that $\sum_{i=1}^{N} \kappa_i = 1$ but otherwise arbitrary.

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - The only valid operator is a weighted mean
  - E.g., median is not valid for $V = V_1 + V_2$

V	V_1	V_2
3	1	2
5	0	5
6	2	4
5	1	4

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - The only valid operator is a weighted mean
  - So the arithmetic mean is valid for $V = V_1 + V_2$ (i.e., WM with $\kappa_i = 1/3$)

V	V_1	V_2
3	1	2
5	0	5
6	2	4
14/3	1	11/3

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - The number of elements in each partition element is not known
  - So, it is difficult to define *a priori* weights κ_i
  - In addition, the order of the elements should be irrelevant

- Proposition 3.
  - If we add symmetry:
    $$\mathbb{C}(x_1, \ldots, x_N) = \mathbb{C}(x_{\pi(1)}, \ldots, x_{\pi(N)})$$
    for an arbitrary permutation π, then the most general solution is
    $$\mathbb{C}(x_1, \ldots, x_N) = (1/N) \sum_{i=1}^{N} x_i$$

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - The number of elements in each partition element is not known
  - So, it is difficult to define *a priori* weights κ_i
  - In addition, the order of the elements should be irrelevant
- An alternative: if $x_1 = x_2$, define $\kappa(x_1) = \kappa(x_2)$
  - According to Prop. 1, κ should be the same for all variables
  - The approach in most clustering algorithms follows this approach
  - E.g. in Fuzzy c-means for records $x_1, \ldots, x_N$ with memberships to the cluster equal to $\mu_1, \ldots, \mu_N$, $\rightarrow$ define

    $$\kappa_i = \frac{(\mu_i)^m}{\sum_{k=1}^{n} (\mu_k)^m}$$

    and then use the function $\mathbb{C}$.
  - This definition satisfies Prop. 1

Microaggregation and Edit Constraints

Nonlinear Constraints

Microaggregation and the edit constraints

- Microaggregation and nonlinear constraints:
  - We apply a similar approach:

V	V_1	$\dots$	V_K
x_1	$x_{1,1}$	$\dots$	$x_{1,K}$
$\vdots$	$\vdots$		$\vdots$
x_N	$x_{N,1}$	$\dots$	$x_{N,K}$
$\mathbb{C}(x_1, \dots, x_N)$	$\mathbb{C}(x_{1,1}, \dots, x_{N,1})$	$\dots$	$\mathbb{C}(x_{1,K}, \dots, x_{N,K})$

  - Now,
$$\mathbb{C}(x_1, \dots, x_N) = \prod_{i=1}^{K} \mathbb{C}(x_{1,i}, \dots, x_{N,i})^{\alpha_i}$$
  - If the original data satisfy this constraint (i.e., $x_j = \prod_{i=1}^{N} x_{j,i}^{\alpha_i}$),
$$\mathbb{C}\left(\prod_{i=1}^{K} x_{1,i}^{\alpha_i}, \dots, \prod_{i=1}^{K} x_{N,i}^{\alpha_i}\right) = \prod_{i=1}^{K} \mathbb{C}(x_{1,i}, \dots, x_{N,i})^{\alpha_i}$$

Microaggregation and the edit constraints

- Microaggregation and nonlinear constraints:
  - Proposition 4.

    $\mathbb{C}$ a function satisfying

    $$\mathbb{C}(\prod_{i=1}^{K} x_{1,i}^{\alpha_i}, \ldots, \prod_{i=1}^{K} x_{N,i}^{\alpha_i}) = \prod_{i=1}^{K} \mathbb{C}(x_{1,i}, \ldots, x_{N,i})^{\alpha_i}$$

    for given values $\alpha_1, \ldots, \alpha_K$ $(\alpha_i \neq 0)$ and arbitrary values $x_{i,j}$ for $1 \leq i \leq N$ and $1 \leq j \leq K$, and reflexivity

    $$\mathbb{C}(x, \ldots, x) = x$$

    Then, the most general solution for $\mathbb{C}$ is a function of the form

    $$\mathbb{C}(x_1, \ldots, x_N) = \prod_{i=1}^{N} x_i^{\kappa_i}$$

    for κ_i such that $\sum_{i=1}^{N} \kappa_i = 1$ but otherwise arbitrary.

Microaggregation and the edit constraints

- Microaggregation and nonlinear constraints:
  - Results similar to the linear case (Propositions 5 and 6):
    * Same function $\mathbb{C}$ when arbitrary $\alpha_1, \ldots, \alpha_K$
    * Equal weights when symmetry is added:

$$\mathbb{C}(x_1, \ldots, x_N) = \prod_{i=1}^{N} x_i^{1/N}$$

Microaggregation and Edit Constraints

Constraints on the Values

Microaggregation and the edit constraints

- Linear constraints, and constraints on the values
  - Simple formulation: data define an interval
    - Cluster representative in the interval defined between the minimum and the maximum of the elements in the cluster (internality).

$$\min x_i \leq \mathbb{C}(x_1, \ldots, x_N) \leq \mathrm{max}_i$$

  - Proposition 7. Adding internality to Proposition 1:

$$\mathbb{C}(x_1, \ldots, x_N) = \sum_{i=1}^{N} \kappa_i x_i$$

    for κ_i such that $\sum_{i=1}^{N} \kappa_i = 1$ and $\kappa_i \geq 0$ but otherwise arbitrary.

Microaggregation and the edit constraints

- Nonlinear constraints, and constraints on the values
  - Simple formulation: data define an interval
    - cluster representative in the interval defined between the minimum and the maximum of the elements in the cluster (internality).

$$\min x_i \leq \mathbb{C}(x_1, \ldots, x_N) \leq \mathrm{max}_i$$

  - Proposition 8. Adding internality to Proposition 4:

$$\mathbb{C}(x_1, \ldots, x_N) = \prod_{i=1}^{N} x_i^{\kappa_i}$$

    for κ_i such that $\sum_{i=1}^{N} \kappa_i = 1$ and $\kappa_i \geq 0$ but otherwise arbitrary.

Microaggregation and Edit Constraints

One variable governs the possible values of another variable

Microaggregation and the edit constraints

- One variable governs another one
  - We cannot constraint microaggregation so easily in this case.
  - Study in a case by case basis.
  - Examples (from 1st section):
    EC-GV1: If *sex=male* THEN *number of pregnacies = 0*
    EC-GV2: IF *age* < 17 THEN *gross income* $<$ *mean income*
    EC-GV3: *harvested acres* $\leq$ *planted acres*

Microaggregation and the edit constraints

- One variable governs another one
  - Study in a case by case basis: Case EC-GV3
    EC-GV3: *harvested acres* $\leq$ *planted acres*
  - General case for variables V_1 and V_2 ($V_1 \leq V_2$):

V_1	V_2	$\cdots$	V_K
$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,K}$
$\vdots$	$\vdots$		$\vdots$
$x_{N,1}$	$x_{N,2}$	$\cdots$	$x_{N,K}$
$\mathbb{C}(x_{1,1}, \ldots, x_{N,1})$	$\mathbb{C}(x_{1,2}, \ldots, x_{N,2})$	$\cdots$	$\mathbb{C}(x_{1,K}, \ldots, x_{N,K})$

  - Assumptions and results ...

Microaggregation and the edit constraints

- One variable governs another one
  - General case for variables V_1 and V_2 ($V_1 \leq V_2$):

V_1	V_2	$\cdots$	V_K
$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,K}$
$\vdots$	$\vdots$		$\vdots$
$x_{N,1}$	$x_{N,2}$	$\cdots$	$x_{N,K}$
$\mathbb{C}(x_{1,1}, \ldots, x_{N,1})$	$\mathbb{C}(x_{1,2}, \ldots, x_{N,2})$	$\cdots$	$\mathbb{C}(x_{1,K}, \ldots, x_{N,K})$

  - a) We assume that V_1 and V_2 are microaggregated together.
  - b) If data has already been edited,

$$x_{i,1} \leq x_{i,2} \text{ for all records } i$$

  - c) So, the condition can be formalized as:

    if $x_{i,1} \leq x_{i,2}$ for all records i, then

$$\mathbb{C}(x_{1,1}, \ldots, x_{N,1}) \leq \mathbb{C}(x_{1,2}, \ldots, x_{N,2})$$

    That is, $\mathbb{C}$ is monotonic.

Microaggregation and the edit constraints

- One variable governs another one. Results:
  - a) We assume that V_1 and V_2 are microaggregated together.
  - b) If data has already been edited,

    $$x_{i,1} \leq x_{i,2} \text{ for all records } i$$

  - c) So, the condition can be formalized as:

    if $x_{i,1} \leq x_{i,2}$ for all records i, then

    $$\mathbb{C}(x_{1,1}, \ldots, x_{N,1}) \leq \mathbb{C}(x_{1,2}, \ldots, x_{N,2})$$

    That is, $\mathbb{C}$ is monotonic.

- $\mathbb{C}$ in Prop. 3, 6, 7, 8 are monotonic. So, appropriate here.

- Proposition (solutions) (and the particular cases: $\kappa_i = 1/N$):
  - $\mathbb{C}(x_1, \ldots, x_N) = \sum_{i=1}^{N} \kappa_i x_i$
  - $\mathbb{C}(x_1, \ldots, x_N) = \prod_{i=1}^{N} x_i^{\kappa_i}$

    for κ_i such that $\sum_{i=1}^{N} \kappa_i = 1$ and $\kappa_i \geq 0$

Microaggregation and the edit constraints

- One variable governs another one
  - Study in a case by case basis: Case EC-GV1 and EC-GV2
    EC-GV1: If *sex=male* THEN *number of pregnacies = 0*
    EC-GV2: IF *age* < 17 THEN *gross income* $<$ *mean income*
  - Partition the file (horizontally) and microaggregate each subset[5].
    EC-GV1: Partition $X = \{\Pi_1, \Pi_2\}$,
    Π_1 with *sex=male* and Π_2 with *sex=female*.
    $\rightarrow$ any function $\mathbb{C}$ s.t. $\mathbb{C}(0, \ldots, 0) = 0$ is appropriate
    EC-GV2: Partition $X = \{\Pi_1, \Pi_2\}$,
    Π_1 with *age* < 17 and Π_2 with *age* ≥ 17.
    $\rightarrow$ any monotonic function $\mathbb{C}$ is appropriate

[5] Similar to: Shlomo, N., De Waal, T. (2008), Protection of micro-data subject to edit constraints against statistical disclosure, Journal of Official Statistics 24:2 229-253.

Microaggregation and Edit Constraints

Values are restricted to exist in the domain

Microaggregation and the edit constraints

- Values are restricted to exist in the domain
  - In previous propositions,
    only possible when $\kappa_i = 1$ for a particular i.
  - In general,
    adding this constraint to previous propositions results into:
    a overconstrained problem
    $\rightarrow$ i.e., no solution exists
  - Considering this constraint but not the other,
    any order statistic as e.g. the median[6], or boolean max-min functions.

[6] as used in: Sande, G. (2002) Exact and approximate methods for data directed microaggregation in one or more dimensions, Int. J. of Unc., Fuzz. and Knowledge Based Systems 10:5 459-476.

Conclusions

Conclusions

- Microaggregation is specially suited when constraints are considered
- Analysis of the approaches when defining the centroids