# General Discrete-data Modeling Methods for Producing Synthetic Data with Reduced Re-identification Risk that Preserve Analytic Properties

william.e.winkler@census.gov

Outline

1. Background on Modeling/Edit/Imputation
2. General Fitting/Modeling Procedure
3. Data
4. Results
5. Discussion
6. Concluding Remarks

**Background on Modeling/Edit/Imputation for Discrete Data**

Generalized, parameter-driven methods suitable for use in many different surveys

Based on model of Fellegi and Holt (*JASA* 1976)

*Principles*

1. The minimum number of fields in each edit-failing record $r_0$ should be changed to create an edit-passing record $r_1$ (*error localization*).
2. Imputation rules should be derived automatically from the edit rules.
3. When imputation is necessary, it should maintain marginal and joint distributions of fields.

Table. Two Explicit edits and One Implicit Edit

| | |
|---|---|
| $E^1$ = {*age*<16, *married*, } | |
| $E^2$ = { , *not married*, *spouse*} | $E^1$ & $E^2 \Rightarrow E^3$ |
| $E^3$ = {*age*≥16 , , *spouse*} | |

$r_0$ = {*age*<16, *not married*, *spouse*}. Assume $E^3$ not available.

$r_0$ fails $E^2$. Change Marital_Status to *married* to obtain $r_1$.
$r_1$ now fails $E^1$

If $E^3$ available then change Relat-HH to some other value to obtain record $r_2$ that fails no edits.

## Obvious difficulties

1. *Very*, *very* difficult to effectively enumerate all of the if-then-else rules (large surveys 100s rules).
2. Typically no effective test decks.
3. Logical errors in the entire set of if-then-else rules (and errors in the computer code).

4. After passing through edit software, *no assurance* that a record $r_0$ would be corrected to a record $r_1$ that no longer failed edits.

Operations Research skills (set-covering algorithms, integer programming) needed to implement

Current implementations (late 1990s to present)

- GEIS-> Banff (Statistics Canada economic)
- LEO Statistics Netherlands (both continuous and discrete)
- SCIA Italy (discrete demographic)
- DIA Spain (discrete demographic)
- SPEER, DISCRETE (Census economic, demographic, respectively)

Check logical consistency of system

Find EL (error localization solution)

- Relatively straightforward to apply.

*Do not impute according to any principled methods/models*.

Winkler (2003) connected FH editing with imputation as in Little and Rubin (2002, Chapter 13)

Winkler (2008) created fast generalized software for modeling/edit/imputation and *production*. Software suitably fast for all surveys. Demonstrated how methods are much easier to apply and how exceptionally poorly *well-implemented* hot-deck-based methods were.

Winkler (2008) also showed how to scale microdata to external benchmark constraints using convex constraints.

Generalized software modules

1. module to find all edits (structural zeros)
2. modeling module (iterative fitting)
3. error localization and imputation

Generalized EM Algorithm to produce model

Data $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ represents *n* fields.

Each field $X_j$ takes values $x_{ij}$, $1 \le i \le n_j$.

Each record $r_0$ in **X** has the form

$$\{ X_1 = x_{j1,1}, X_2 = x_{j2,2}, \cdots, X_n = x_{jn,n} \}$$

We may have N records (rows) in **X**. There are *npat* = $n_1 n_2 \ldots n_n$ data patterns. Typically, N<<*npat*. The probability of a given pattern is given by

$$P( X_1 = x_{j1,1}, X_2 = x_{j2,2}, \cdots, X_n = x_{jn,n}) = p_k , \qquad (1)$$

$k \in I$ where $| I | =$ *npat*.

## Convex constraint

$J \subset I$ = set of cells, $p_k$ probability in cell k, $c_k > 0$

s.t. $\Sigma_{k \in J} c_k = 1$, $b > 0$

Convex constraint: $\Sigma_{k \in J} c_k p_k \leq b$

General justification for convex constraints (also structural zeros) in an iterative fitting procedure (Winkler 1990, *Ann. Prob.*)
General EMH procedure (Winkler 1993) under convex constraints that generalizes the MCECM procedure of Meng and Rubin (1993)

## Reduce Re-identification Risk

Convex constraints can be used to put lower and upper bounds on individual cell probabilities
Preserve Analytic Properties (also adjust to benchmark constraints)
Put lower and upper bounds on margins.

## Data from UCI machine learning repository 'Adult'

Initially used: WorkClass (7 values), MaritalStatus (7 values),
Race (5 values), and Sex (2 values)

490 ($7 \times 7 \times 5 \times 2$) data patterns, 45221 data records

80 small cells with count 1 or 2 (count of 103 records)
191 cells that are sampling zeros
290 cells having count above 2

EM Procedure – Fit count 103 over 271 (80+191) cells
Starting value 103/271 in each cell

Determine all 3-way interaction fits best
Repeat fitting with upper bound 0.000004 on all small cells

Overall Fit (epsilon 0.00000000001)

| | |
|---|---|
| Maximum Likelihood | -3.234682 |
| Likelihood Linear | -3.234982 |
| Likelihood Convex | -3.241030 |

Perform Fitting -> Model

With model, randomly draw probability proportional to size until reach 45221 records

Can create multiple copies of data

Table 1. Original and Fitted Probabilities for Selected Cells

| Cell | Original | Fitted |
|---|---|---|
| 0000 0 0 0 0 | 0.02859 | 0.02876 |
| 0001 0 0 0 1 | 0.25344 | 0.25328 |
| 0002 0 0 1 0 | 0.00172 | 0.00163 |
| 0003 0 0 1 1 | 0.00781 | 0.00790 |
| 0004 0 0 2 0 | 0.00031 | 0.00037 |
| 0005 0 0 2 1 | 0.00181 | 0.00175 |
| 0006 0 0 3 0 | 0.00042 | 0.00042 |
| 0007 0 0 3 1 | 0.00210 | 0.00210 |
| 0020 0 2 0 0 | 0.09670 | 0.09636 |
| 0021 0 2 0 1 | 0.12426 | 0.12460 |
| 0107 1 3 3 1 | 0.00002 | 0.00000 |
| 0301 4 2 0 1 | 0.00637 | 0.00610 |
| 0487 6 6 3 1 | 0.00000 | 0.00001 |
| 0488 6 6 4 0 | 0.00000 | 0.00001 |
| 0489 6 6 4 1 | 0.00000 | 0.00001 |

Largest deviation cells 0020, 0021, 0301

Table 2. Original and Fitted 3-way Margins
for Selected Marginal Cells

| Pattern = 3, Variables 1,3,4 | | |
|---|---|---|
| 00000 | 0.205988 | 0.205988 |
| 00001 | 0.427102 | 0.427102 |
| 00002 | 0.007607 | 0.007589 |
| 00003 | 0.013511 | 0.013518 |
| 00004 | 0.002211 | 0.002223 |
| 00005 | 0.003936 | 0.003925 |
| 00006 | 0.002410 | 0.002423 |
| 00007 | 0.004179 | 0.004146 |
| 00014 | 0.000133 | 0.000028 |
| 00054 | 0.000199 | 0.000099 |
| 00067 | 0.000000 | 0.000210 |

Largest deviations at cells
0067, 0014, 0054

6 Variable scenario
588,160 ($74 \times 7 \times 7 \times 16 \times 5 \times 2$) data patterns
9447 cells having count 1 or 2
3098 cells having count above 2
~98% cells are sampling zeros

More flexibility in assigning positive probability to originally sampling zero cells in order to preserve analytic properties

Draw 1-3 copies of 45221 records from resultant model.

For both 4- and 6-variable scenarios

*Cannot re-identify using record linkage*.

Data are neither epsilon-private nor epsilon-delta private

## Discussion

Epsilon-private methods have *not been shown to preserve analytic properties* (Dwork 2006, Barak et al. 2007, Dwork et al. 2007a and Dwork and Yekhanin 2008).

Partial exception is based on the epsilon-delta-privacy ideas of Machanavajjhala et al. (2008, ICDE). Narrowly focused 'on-the-map' application.

*Computational issues* (Xiao and Tao 2008 VLDB) related to computing $L^1$ – sensitivity of functions needed for justifying epsilon-privacy.

*Interesting Experiment*: Produce 3 synthetic copies of 588,160-size microdata associated with 6 variables. See if the original data can be reconstructed using cryptographic methods.

## Concluding Remarks

Although modeling/iterative-fitting tools are flexible and sufficiently fast for *all* edit/imputation applications, the methods need to be enhanced and speeded-up for confidentiality/privacy applications.