Optimal Microdata File Merging:
A New Model & Network Optimization Algorithm

Richard S. Barr, *SMU*
Betty L. Hickman, *U Nebraska Omaha*
J. Scott Turner, *Oklahoma State*

Microdata Files

- Stratified samples of large populations
- Multi-attribute representations of the underlying distributions and interactions
- Expensive to create

Example Microdata

- Statistics of Income
- Current Population Survey
- American Housing Survey
- Census of Agriculture
- Decennial Census
- Economic Census
- Integrated International Census
- Canadian Families
- Survey of Income and Program Participation
- Commodity Flow
- Foreign Trade
- County Business Patterns
- Population & Housing
- Nat'l Survey of Fishing, Hunting & Wildlife
- National Health Interview

Limitations of Individual Samples

- Data are often required that are
  - Not part of the current source
  - Of superior quality

- Choices available:
  - Commission a new study
  - Ignore variables
  - Impute missing items
  - Merge two files to combine surveys

Merging Microdata Files

- Two microdata samples
  - Are drawn from the same population
  - Include record *weights*, reflecting sampling rate
    - A record weight of 10 reflects 1:10 sampling rate
    - Record represents 10 population units
- Files A and B are merged to form file C
  - Composite C has data items from both A and B
  - A-B record pairs are matched, based on common attributes

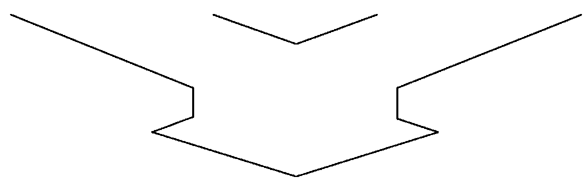
Microdata Merge Diagram

FILE A RECORDS

2000	$AGI1_1$ $STATE1_1$	CAP $GAIN1_1$
500	$AGI1_2$ $STATE1_2$	CAP $GAIN1_2$
⋮	⋮	⋮
a_i	$X1_{1i} \ldots X1_{R1}$	$Y_{1i} \ldots Y_{Si}$
RECORD WEIGHT	COMMON ITEMS	FILE A ONLY

FILE B RECORDS

500	$AGI2_1$ $STATE2_1$	CAP $GAIN1_1$
500	$AGI2_2$ $STATE2_2$	CAP $GAIN1_1$
⋮	⋮	⋮
b_j	$X2_{1j} \ldots X2_{Rj}$	$Z_{1j} \ldots Z_{Tj}$
RECORD WEIGHT	COMMON ITEMS	FILE B ONLY



FILE C (COMPOSITE RECORDS)

x_{ij}	$X1_{1i} \ldots X1_{R1}, Y_{1i} \ldots Y_{Si}$	$X2_{1j} \ldots X2_{Rj}, Z_{1j} \ldots Z_{Tj}$

INTERRECORD DISSIMILARITY MEASURE (DISTANCE FUNCTION)

$$c_{ij} = f(X1_{1i}, \ldots, Y_{Si}, X2_{1j}, \ldots, Z_{Tj})$$

Microdata Matching Methods

- *Exact matching* uses unique-valued common items
- *Statistical matching* or *merging*
  - Mates similar records
  - Using non-unique common items
- Exact matches are
  - Always preferable
  - Rarely possible or permitted by law

Statistical Merging Techniques

- *Unconstrained* merges
  - Use a base file (A) and augmentation file (B)
  - Each base-file record matched with "most similar" file B record
    - Matching with replacement
    - Ignores file B's record weights
  - Greatly distorts the statistical characteristics of B's items

Statistical Merging Techniques

- Constrained merges
  - Weight constraints added to ensure records in each file are not over- or under-matched
    - The sum of each record's matched weights = original weight
    - One record may be matched with multiple records in the other file
    - Matching without replacement
  - Statistical characteristics of both file's values are preserved

Constrained File-Merge Model

- Given:
  - A_i= record i weight in file A
  - B_j = record j weight in file B
- Assumed: equal population size:

$$\sum_i A_i = \sum_j B_j$$

- Weight constraints

$$\sum_j w_{ij} = A_i, \text{ for all } i$$

$$\sum_i w_{ij} = B_j, \text{ for all } j$$

$$w_{ij} \geq 0, \text{ for all } i, j$$

where w_{ij} = weight of composite record (i,j)

Optimal Constrained Merge

Minimize $\sum_i \sum_j c_{ij} w_{ij}$

subject to: $\sum_{j=1}^{n} w_{ij} = A_i, \quad \text{for all } i$

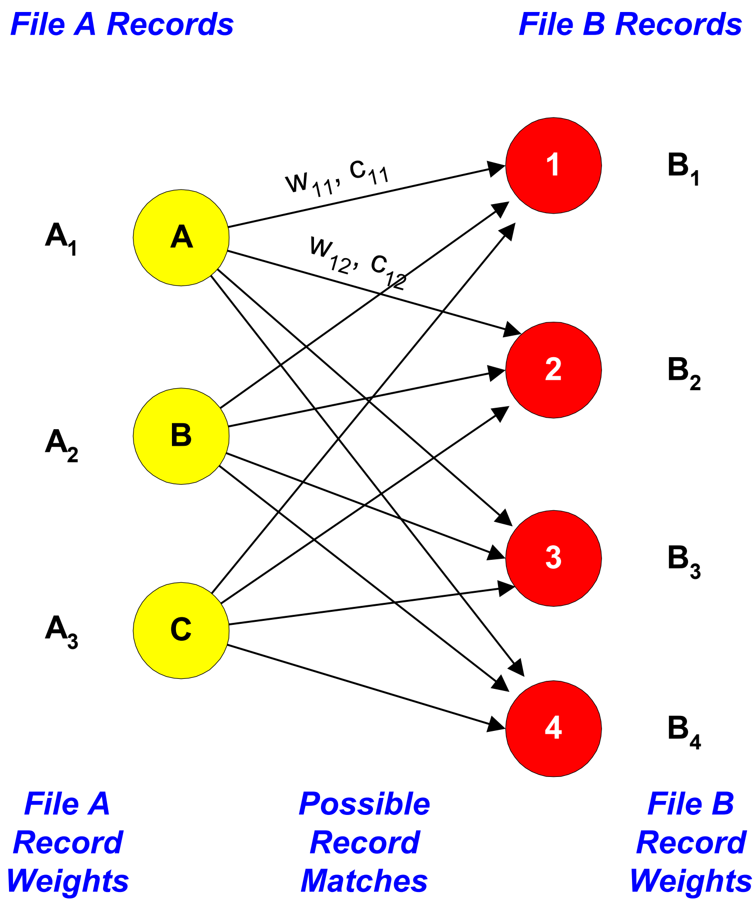
$$\sum_{i=1}^{m} w_{ij} = B_j, \quad \text{for all } j$$

$$w_{ij} \geq 0, \qquad \text{for all } i, j$$

where c_{ij} = dissimilarity measure (distance) between record i in A and j in B (Turner and Robbins, U.S. Treasury)

Optimal Constrained Merge Model

- Has the form of a transportation problem
- One source node for each file A record
- One sink node for each file B record
- One arc for each record-match possibility



Problem Characteristics

- Large network models
  - 1,000s of nodes (constraints)
  - Millions of variables (*mn* arcs)
- U.S. Treasury: Optimal merge system
  - In use since mid-1970s
  - Described in Barr and Turner, 1980
  - Routinely solves problems with 20,000 constraints and 30-million variables

Underlying Rationale for Merging

Assumptions

- File A = $\{X_1,Y\}$ and file B = $\{X_2,Z\}$, where
  - X_1, X_2 = common items
  - Y = set of data items found only in file A
  - Z = items found only in B
- Both A and B are valid samples from the same population

Objectives of Merging

- Form a sample file $C = \{X_1, X_2, Y, Z\}$
  - Such that C corresponds statistically to a $\{X, Y, Z\}$ sample taken from the same population
- Make inferences about (Y,Z) and $(Y,Z|X)$ relationships using C
  - Can already infer (X,X), (X,Y), and (X,Z)

Problems and Criticisms

- Conditional independence assumption, CIA
  - If YZ relationships left out of merge process, merged files tend to yield correlations $r_{Y,Z} \approx 0$
- Effect of $r_{Y,Z} \approx 0$
  - Depends on usage of the file
  - From negligible to disturbing

New Optimal Merge Model

- Incorporates outside information
- YZ relationships included in the model via
  - Penalties for illogical combinations
  - Estimates of second-order information
    - Based on intermittent samples, logic, or best guesses
    - For cases where CIA is unreasonable

Covariance $s_{y,z}$ Computation

Y-Z covariance:

$$s_{y,z} = \frac{1}{W} \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} (y_i - \overline{y})(z_j - \overline{z})$$

where

W = sum of record weights

y_i = value of item y in ith record of file A

z_j = value of item z in jth record of file B

$\overline{y}, \overline{z}$ = sample means for y and z

Correlation Computation

Y-Z correlation:

$$r_{y,z} = \frac{s_{y,z}}{\sigma_y \sigma_z}$$

where

σ_y, σ_z = standard deviations of y in A and z in B

Including a Correlation Constraint

- If an estimate for the correlation parameter is ρ, a goal-programming side condition is:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij} w_{ij} \approx \rho$$

where

$$d_{ij} = \frac{(y_i - \overline{y})(z_j - \overline{z})}{W\sigma_y\sigma_z}$$

Extended Merge Model

$$\text{Minimize } \sum_i \sum_j c_{ij} w_{ij}$$

$$\text{subject to: } \sum_j w_{ij} = A_i, \qquad \forall i$$

$$\sum_i w_{ij} = B_j, \qquad \forall j$$

$$\sum_i \sum_j d_{ij}^k w_{ij} \approx \rho_k, \qquad \forall k \text{ parameters}$$

$$w_{ij} \geq 0, \qquad \forall i, j$$

Model Characteristics

- A network with “side conditions”
- Network component:
  - Large, dense
  - Must be feasible
- Side conditions:
  - Few to many
  - Dense LHS, RHS are estimates & targets
  - Feasibility desirable, may not be possible

Network with Side Conditions Algorithm

NSC Problem

$$\begin{array}{llll} P: & \text{Min} & cx = z & \\ & \text{s.t.} & Ax = b & \textit{Network constraints} \\ & & Dx = d \pm \varepsilon & \textit{Side conditions} \\ & & x \geq 0 & \end{array}$$

Lagrangean Approach

Dualize the side conditions, ignoring error

$$
\begin{aligned}
LR(\lambda):\ z_d(\lambda) = \text{Min } & cx + \lambda\,(Dx-d) \\
\text{s.t. } & Ax = b \\
& x \geq 0
\end{aligned}
$$

where λ = vector of Lagrangean multipliers

Subgradient Method

1. Begin with initial multiplier vector λ^0, k=0
2. While x^k is infeasible to P (or other rule):
   - Generate λ^{k+1} using: $\lambda^{k+1} = \lambda^k + t^k(Dx^k\text{-}d)$
     - where x^k is an optimal solution to LR(λ^k), and
     - t^k is a positive scalar step size
   - $k = k + 1$
   - Solve LR(λ^k)

Implementation Characteristics

- Stepsize:

$$t^k = \frac{\theta^k \left| \overline{z} - z_d(\lambda^k) \right|}{\left\| Dx^k - d \right\|^2}$$

where $0 < \theta < 2$, $\overline{z} = z_d(0)$, an estimate of z

- Stopping criteria
  - Within ε-tolerances and 10% of z(0)
  - Cost of solution cx^k unchanged in p iterations
  - Iteration limit exceeded

Empirical Analysis

- Randomly generated multi-normal test data
- XYZ datasets with predetermined correlations were generated
- Records were divided into
  - "File A" records, with Z values removed
  - "File B" records, with Y values removed
- Attempted to construct File C with target correlation values

Test Sets

Test Set:	A	B
File A size:	100, 300	400, 1000
File B size:	200, 300	600, 1000
Correlations:	4 to 25	4 to 25
Possible pairs:	20K, 90K	240K, 1M

Solution Software

- PPNET-SC code
  - Based on parallel network optimizer, PPNET
  - Incorporates side-conditions
- Compared with NETSIDE, networks-with-side-constraints optimizer

Summary

- The new model and algorithm effectively maintains all YZ relationships included in the model
- The convergence is relatively fast
- Improves the quality of the composite files
- Testing on larger problems is next