Mixed Mode Methods in a World of Social Isolates, Pervasive Surveillance, and Ubiquitous Transaction Records: A Modest Proposal

Robert M. Groves and Trivellore Raghunathan
University of Michigan and
Joint Program in Survey Methodology

Outline

1. What are modes and why combine them?
2. Pervasive role of nonresponse and costs
3. Traditional mixed mode paradigms
4. Data assembly versus data collection
5. Technical impediments
6. Social/legal impediments
7. Next steps

What are modes and why combine them?

- Modes are not merely alternative communication media
  - sampling frames
  - recruitment protocols
  - presence of interviewers
  - visual vs. audio vs. physical presence
- Hence, modes inherently vary in
  - costs
  - coverage properties of target population
  - sampling design effects
  - response rate characteristics
  - essential measurement properties

Pervasive Role of Nonresponse

- Most mixed mode designs attempt minimizing cost and nonresponse rates
- Most mixed mode designs hope for absence of measurement error differences
- Nonresponse rates are falling in the rich countries of the world, apparently in all modes

The strongest influences toward mixed mode designs are costs, coverage, and nonresponse issues; measurement generally trails in importance

Example

- Random digit dialed telephone surveys in the US have experienced dramatically increases in nonresponse rates and costs per interview
- Mobile-phone only populations threaten coverage properties of RDD surveys

More and more practitioners are questioning the feasibility of single mode phone designs

Traditional Mixed Mode Paradigms

- Sequential application
  - begin with cheap mode, use more expensive as nonresponse rate reduction
- Multiple frame, multiple mode
  - measure those on cheap frame using cheap method
  - "fill in" noncovered with other frames and modes fitted to the frame
- Respondent driven mixed modes
  - single frame, respondent choice
- Randomized assignment of multiple modes
  - rarely done, but greatly beneficial for estimation

Weakness of Traditional Paradigms

- They either often exist solely to measure mode biases, or
- They often solely exist to obtain data in a cost efficient manner, with little ability to incorporate mode effects into estimation

New Data Opportunities

- Increasingly there exist large data bases containing information on people and their activities
  - commercial credit bureau person records
  - transaction records of customers
  - voting records
  - property records
  - employee records
  - health records

Common Properties of Such Record Systems

- Coverage
  - customers of services or products
- Data content
  - variables relevant to the administration of the service
- Identifying variables
  - names, addresses
  - government identification numbers sometimes

Data Assembly Versus Data Collection

- A new enterprise is developing outside of traditional survey design -- assembling and linking data sets; examples
  - US linking of person surveys with employer surveys
  - US assembly of social security and Medicare data with survey data
  - record matching of national census and other data
  - Germany data fusion efforts
  - commercial efforts at massive matching of data records by name, address, or other varaibles
  - using imputation models in multi-mode settings

Properties of this Data World – A Patchwork of Data

- Large undercoverage of the household population
  - disproportionately transient, young, poor
- Large item missing data rates
  - files distributed with majority of data missing on some variables
- Little concern with measurement properties
- However, massive data bases on increasing numbers of persons

New Approaches for Mixed Mode Stimulated by Combining Survey and Administrative Data

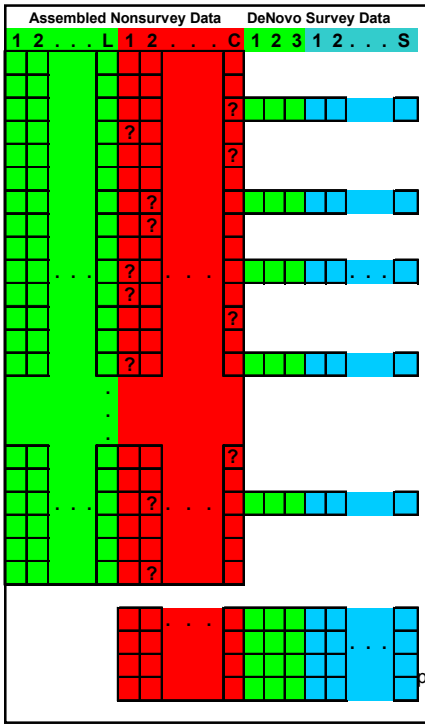
- First, let's begin to think of administrative data sets as a new mode
- Then, let's examine statistical practices and designs used in administrative-survey mixes
- Then, let's ask the question of how survey designers can both exploit these developments and contribute to them

Alternative Designs for Mixed Mode Data Assembly

- Exact match, “fill in” data collection on probability sample
- Exact match, “fill in” data collection on probability sample, imputation on nonsampled cases
- Data fusion, “fill in” data collection, imputation on nonsampled cases
- Randomized mode assignment

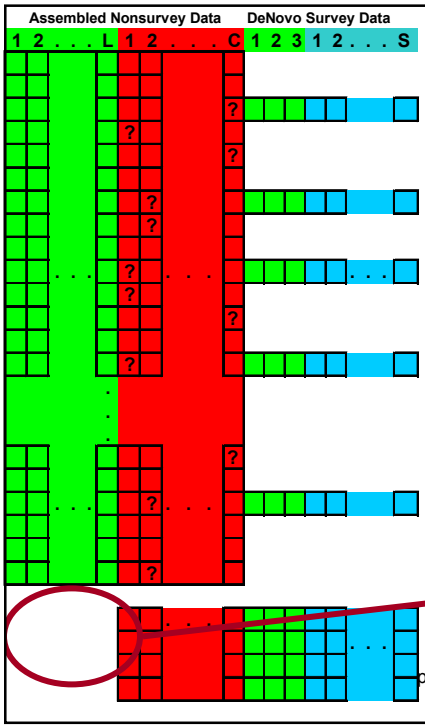
Exact match, “fill in” data collection on probability sample

- Example: link rich administrative frame with inadequate coverage to survey data on other variables
- Example: use one cheap mode on large sample, but expensive second mode of sample of respondents and nonrespondent to first mode



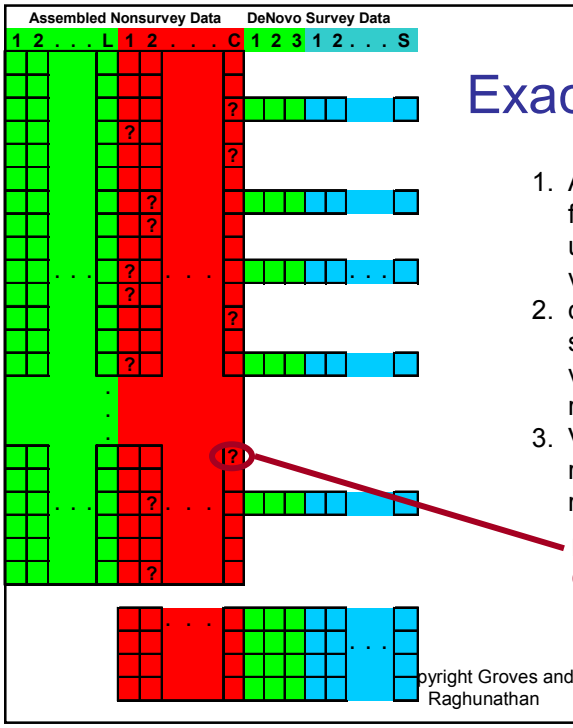
Exact Match, Fill-in

1. Assembled nonsurvey data filled with missing data and unknown coverage errors *but* valuable red variables
2. de novo probability sample survey design to capture valuable blue variables, and red variables where needed
3. Value is reduced burden on respondents for collection of red variables



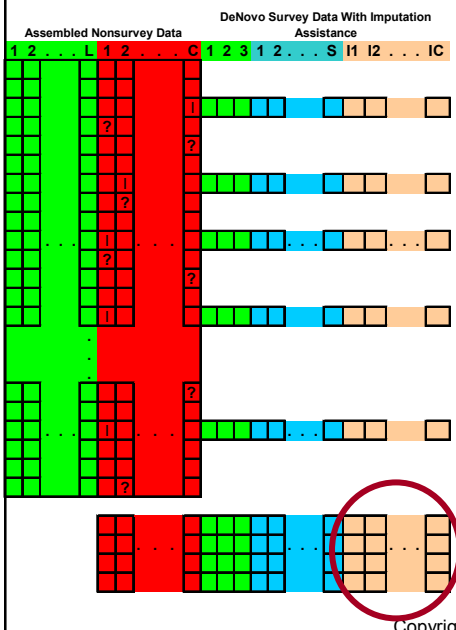
Exact Match, Fill-in

1. Assembled nonsurvey data filled with missing data and unknown coverage errors *but* valuable red variables
2. de novo probability sample survey design to capture valuable blue variables, and red variables where needed
3. Value is reduced burden on respondents for collection of red variables



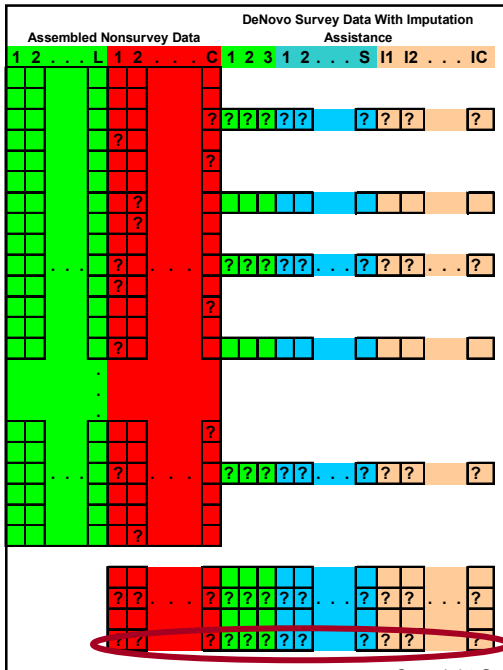
Exact Match, Fill-in

1. Assembled nonsurvey data filled with missing data and unknown coverage errors *but* valuable red variables
2. de novo probability sample survey design to capture valuable blue variables, and red variables where needed
3. Value is reduced burden on respondents for collection of red variables



Exact Match, Fill-in

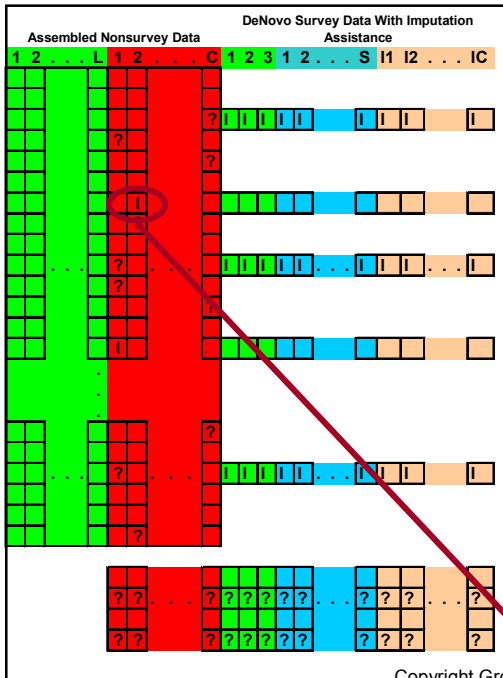
1. **Designer can enhance the accuracy of imputation by collecting good predictors of red variables**
2. **Result is a complete sample survey data set with red and blue variables (note I's among red variables for sampled cases)**
3. **Value is that red variables are available without added respondent burden**



Exact Match, Fill-in, Acknowledging Survey Nonresponse

1. The red variables can be used in imputation of blue variables
2. If pink variables are available on survey nonrespondent cases, they too can be employed in imputation models
3. The sampled survey cases outside the nonsurvey data set, do not have the same predictive set available
4. More realistic outcome of the prior cases

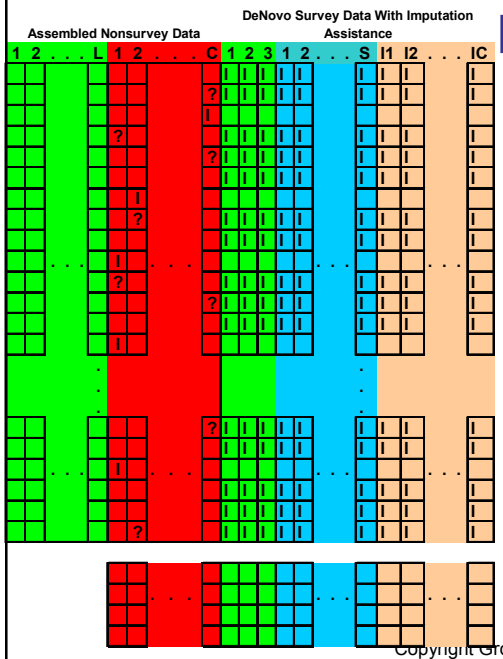
survey unit nonresponse



Exact Match, with Nonsurvey and Survey Data Imputation

1. **Use red variables for imputation of blue and pink variables on survey nonrespondent cases**
2. **Use weighting approach for survey nonresponse from cases outside the nonsurvey data set**
3. **Challenge is whether red variables are predictive of blue variables**
4. **Value is use of red variables for imputation within nonrespondent cases of sample survey**

item imputation



Exact Match, Full Imputation

1. Use relationship between red, blue, and pink variables among survey cases to impute blue and pink variables in nonsampled cases (explicitly MAR)
2. Challenge is whether red variables offer enough predictive power for blue and pink variables
3. Value is nearly complete data set, taking advantage of larger sample size of red variables

Example, Yucel and Zaslavsky (2005)

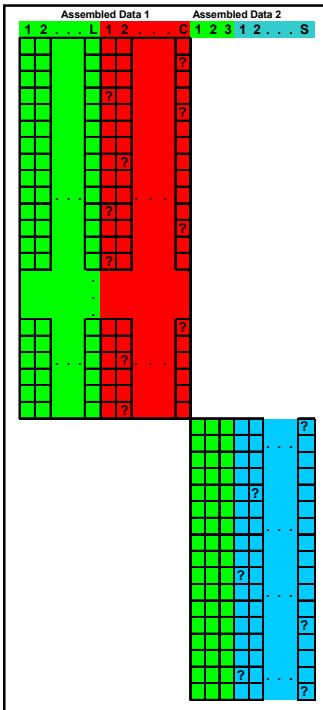
- Cancer registry, with large sample but low quality data
- Followup physician survey on small probability sample but with rich, high quality data
- Use of small survey to model measurement error in larger data set and yield improved estimates on full sample

What Survey Designers Can Do

- imputation for the blue variables needs forethought
- red and pink variables are most useful when they can be measured on probability sample of full target population

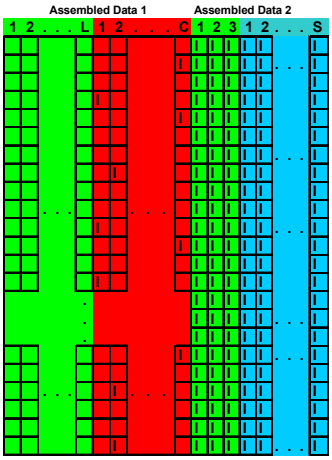
Data Fusion

- Some commercial firms are enhancing data records through statistical matching, not exact matching or through model-based imputation models similar to a statistical matching



Data Fusion Conditions

1. Two data sets with no possibility of exact match
2. One data set has valuable red variables; the other, valuable blue variables

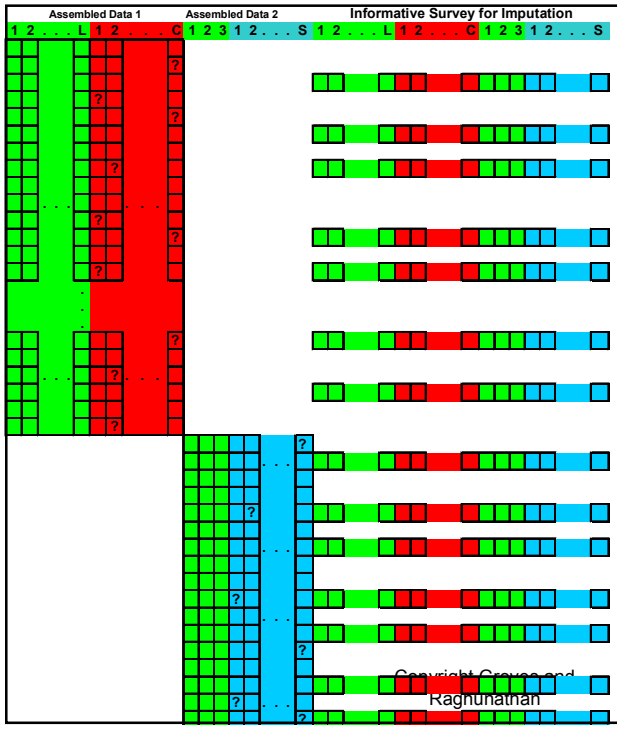


Data Fusion Completed

1. **Statistical matching on green variables, to produce cases containing both red and blue variables**
2. **Value is joint analysis of red and blue variables**
3. **Challenge is quality of statistical match**

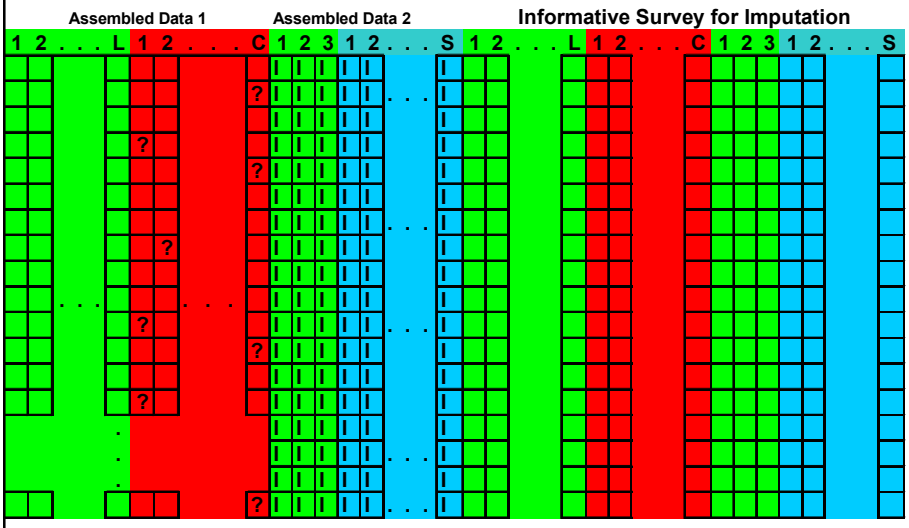
Example: Schenker and Raghunathan (2005)

- Large, high response rate, general purpose health survey
- Small, lower response rate, very rich health survey with physiological measures
- Use the small rich survey to build model of measurement error for self-reports in large survey
- Model used to improve estimates of health condition from large survey



Survey-Assisted Data Fusion

1. Draw probability sample from each data set
2. Collect both red and blue variables on sampled cases
3. Value is measure of mode effect on red and blue variables, to help model measurement error properties

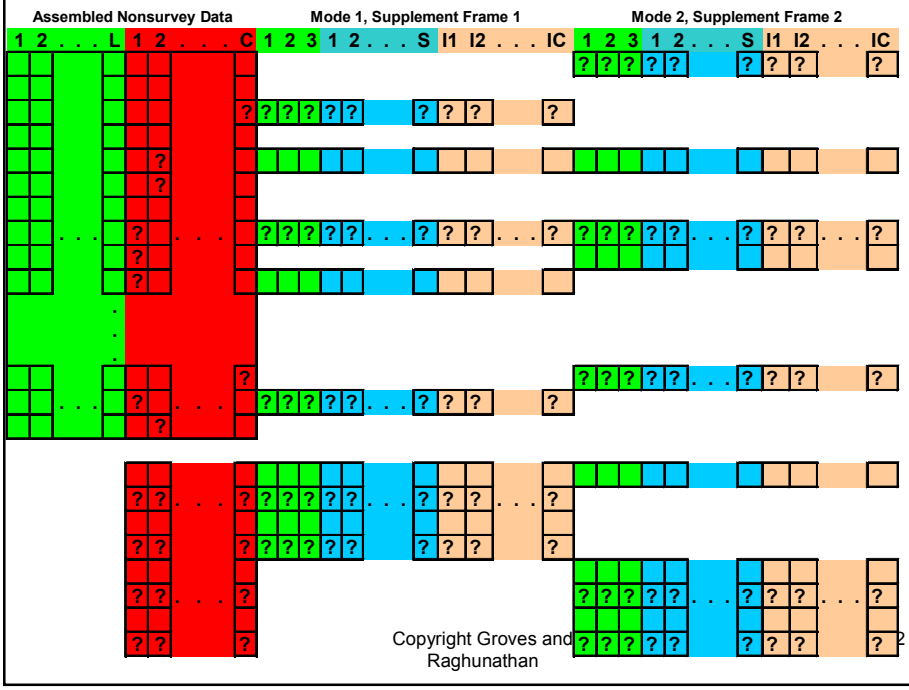


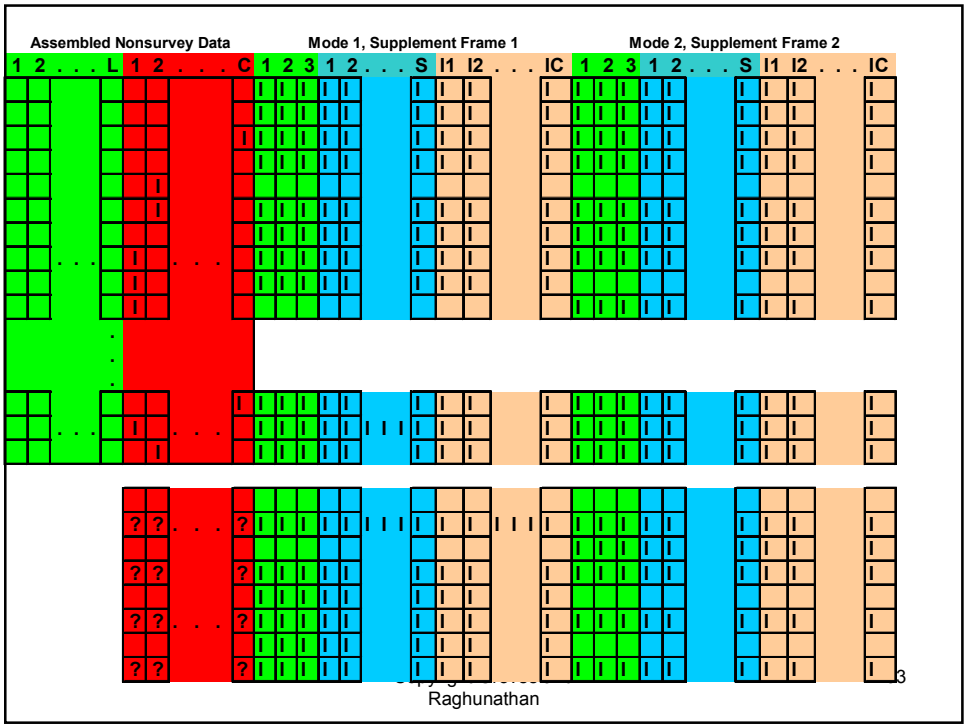
Using the Ideas for Mixed Mode Designs

- multi-mode designs informing
  - coverage errors
  - nonresponse errors
  - measurement errors

Example

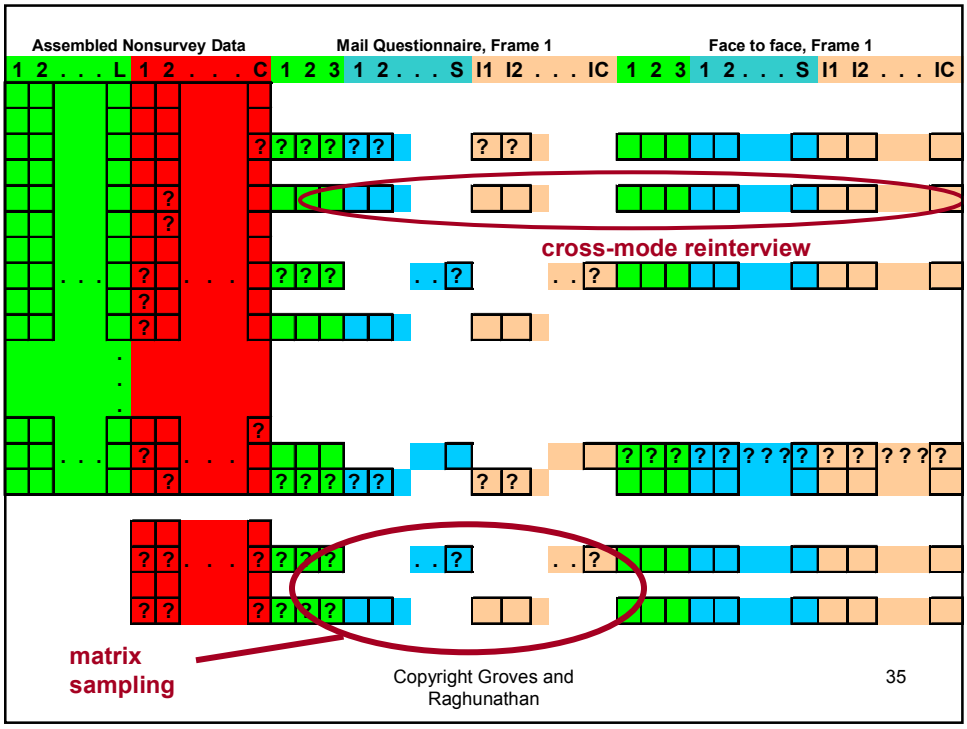
- Acquire administrative record base with, for example, phone numbers and addresses
- Draw repeated samples from record base and supplement samples from number frame and address frame
- Deliberately replicate modes on sample of cases

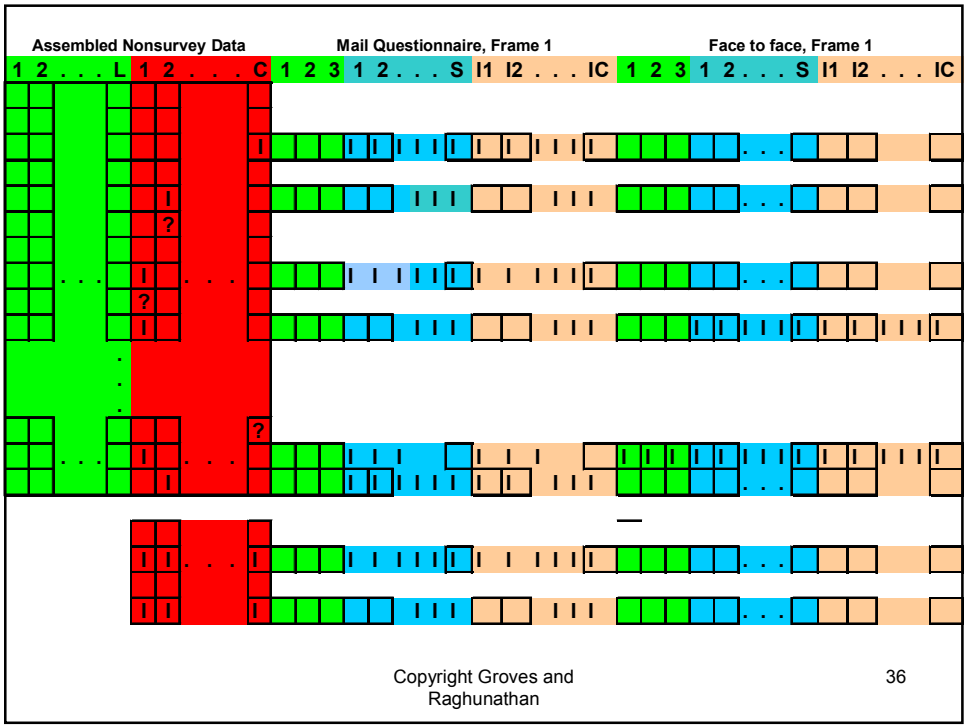




Example

- Combine modes with very different sensitivities to questionnaire length (e.g., mail vs. face to face)
- Implement randomized matrix sampling on mail portion of sample
- Impute for missing data


Assembled Nonsurvey Data
Mail Questionnaire, Frame 1
Face to face, Frame 1
1 2 . . . L 1 2 . . . C 1 2 3 1 2 . . . S I1 I2 . . . IC 1 2 3 1 2 . . . S I1 I2 . . . IC
cross-mode reinterview
matrix sampling
Copyright Groves and Raghunathan
35


Assembled Nonsurvey Data
Mail Questionnaire, Frame 1
Face to face, Frame 1
1 2 . . . L 1 2 . . . C 1 2 3 1 2 . . . S I1 I2 . . . IC 1 2 3 1 2 . . . S I1 I2 . . . IC
Copyright Groves and Raghunathan
36

Features of this Mixed Mode Design

- all cases have imputed or real data for both modes, to measure mode differences in measurement
- some cases have real data on two modes
- nonsurvey data helpful on adjustment of full nonresponse
- use of expanded frame helpful to study coverage errors

Technical Impediments

- Covariance of missingness propensity in existing record systems with survey nonresponse propensity
- Specification of imputation models
- Assessment of models for imputation/fusion

Social/Legal Impediments

- Who “owns” the data?
  - countries differ on rights of persons to control records on themselves
- Under what circumstances will persons agree to give access to their data?
  - can the survey researcher give direct benefits to the respondent in summarizing data?
  - will respondens view requests for access as a burden reduction or threat to privacy?
- Will commercial holders of data permit acquisition for research purposes?
- What societal institutions are necessary for constructing such capabilities?

A Modest Proposal

- A designed, dual frame, mixed mode, matrixed sampled instrument, with imputation
  - base frame -- commercial data base with rich variables, name and address of housholds, some telephone numbers
  - supplement frame –address or person frame
  - mail questionnaire or phone survey, matrixed sampled instrument on full probability sample of commercial data base
    - instrument contains predictors of commerical data base variables
    - all nonrespondents given face to face followup
  - 1/3 sample administered full questionnaire face to face
  - imputation for matrix sampled variates
  - imputation for missing mode variates on 2/3 sample
  - use of base frame for unit nonresponse adjustment

Research Questions

- Levels of coverage error in commerical data set
- Levels of nonresponse error in both modes (using commercial data)
- Variance/bias of imputation for matrix sampled mail questionnaire
- Variance/bias of imputation of missing mode data
- Mode differences in estimates