A Framework for Empirical Cost Modeling Relating Cost & Data Quality

Mary H. Mulry & **Bruce D. Spencer**
U.S. Census Bureau **Northwestern Univ.**

2012 International Total Survey Error Workshop
September 3, 2012

United States™ Census Bureau

Cost Models

- Specify the data quality profiles attainable for a given level of cost (or resources) for a statistical program
- Specify the cost (or resources) required to attain a data quality profile for a statistical program
- Need to be empirically-based rather than theoretically based
- Map costs to data quality profiles for different designs for a statistical program

Empirical data quality (DQ)

- Multi-dimensional for estimate of 1 group
  - Total survey error components: bias, variance
  - Timeliness
- Dimensionality increases with multiple domains, e.g. distributions across domains
  - Race/Hispanic ethnicity groups
  - Subnational geography: states, counties, cities

Empirical cost models

- Formulation requires comprehensive approach
- Factorial experiments may be needed to estimate cost models if interactions are present.
- Example:
  - In a census setting, proposed Operation A & Operation B perform well in independent tests
  - When implemented together, Operation A adds people that the subsequent Operation B deletes

Cost model for statistical program

- Data quality measure (DQ) that can be estimated
- Cost measure (C) that can be estimated
- Then have the pair *(DQ(i), C(i))* for each design *i* under consideration

Goal of cost modeling

- Find a class of designs attaining the optimal DQ profile for each level of cost
  - DQ profiles may be only partially ordered
- Then use cost-benefit analysis to guide selecting a design from the class
  - Consider (quantify if possible) the benefit from each attainable DQ profile
  - Benefit may be multidimensional or too complex to measure in dollars
  - Compare cost versus benefit of data quality

Methodology example

U.S. 2010 Census was multi-mode

- Mailout/mailback questionnaires
- Field follow up of non-responding housing units for personal visit interviews (NRFU)

Research question:

Will completing NRFU with administrative records (AR) data reduce cost and provide acceptable data quality?

Estimate DQ of proposed methods

- Simulations of different methods produce different population totals using
  - 2010 NRFU data
  - Census-like file formed from merging several AR sources
- Estimate variance using replication methods
- To estimate bias, need ‘gold standard’
- Construct ‘gold standard’ using 2010 Census Coverage Measurement Program (CCM) data for sample of blocks

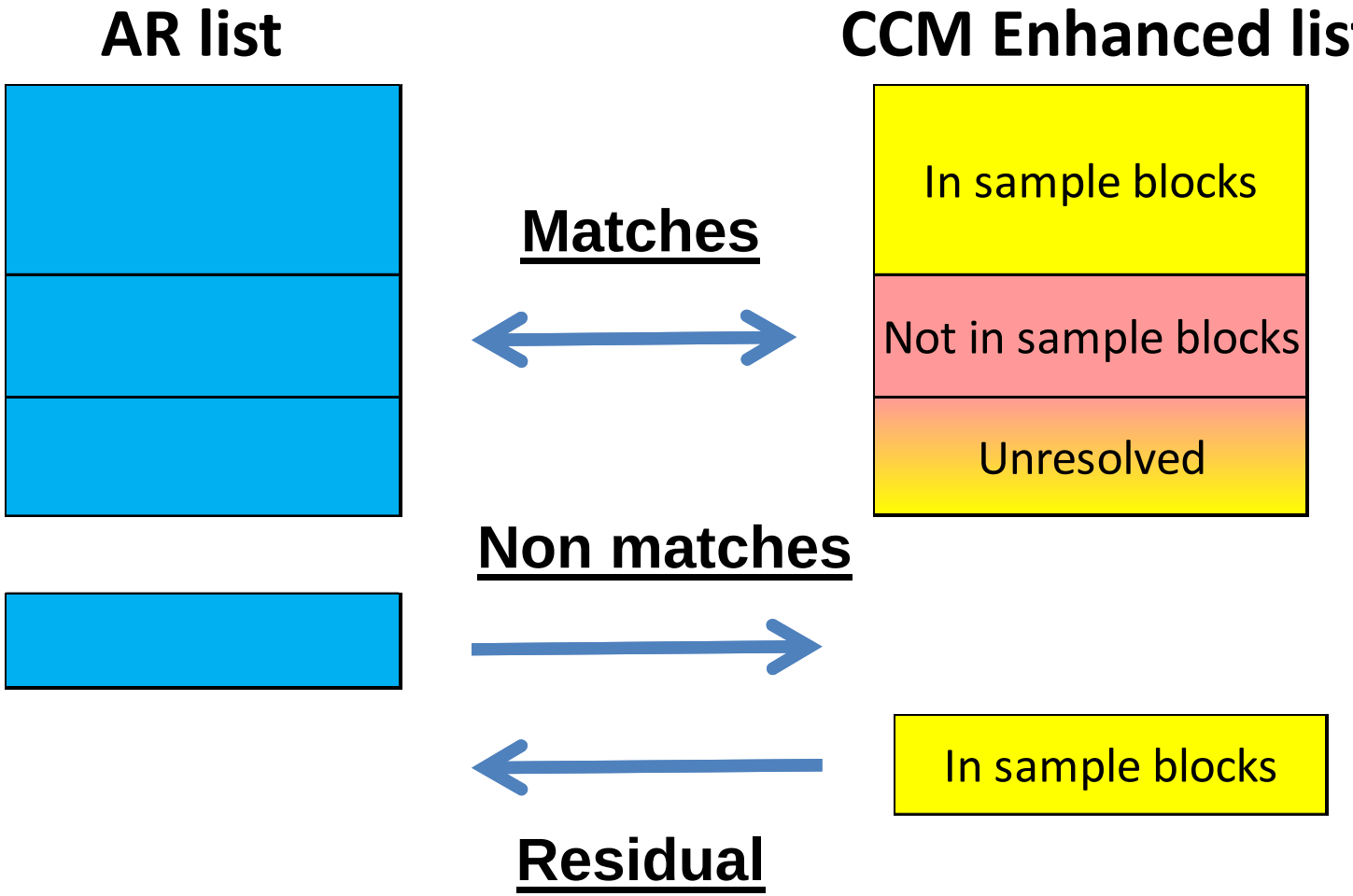
CCM qualifies for ‘gold standard’

- CCM sample data receive intensive evaluation
- Overlapping samples in sample blocks
  - Census enumerations (E) & independent list (P)
- Case-by-case electronic & clerical matching
  - determine accuracy of census enumerations
  - identify people not matching an enumeration
- Nationwide electronic search of census to find duplicate enumerations
- Field follow-up to resolve ambiguities

Merge E & P sample lists to create CCM enhanced list for ‘gold standard’

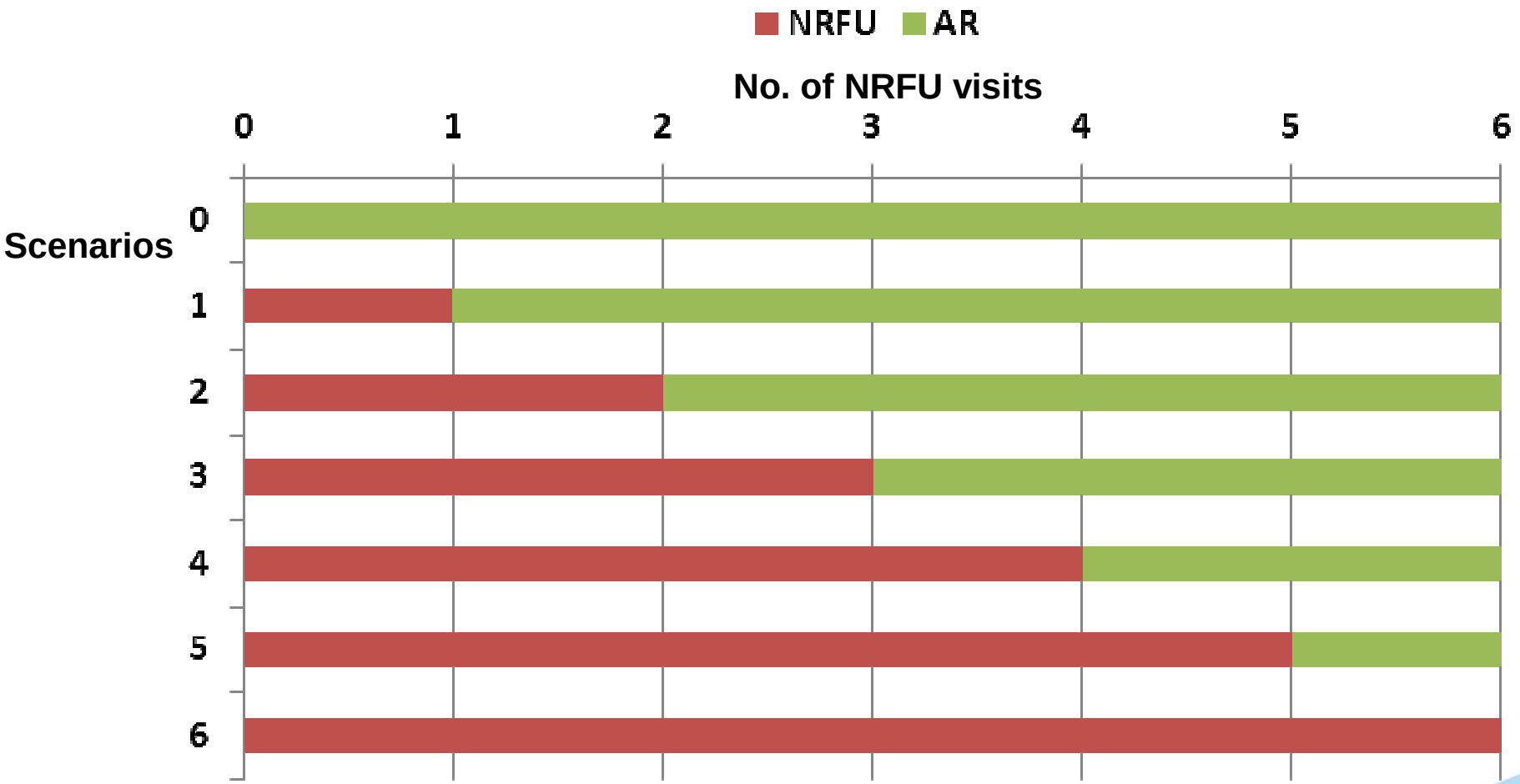
- In sample block on Census Day
  - Correct census enumerations (E sample)
  - People not in the census (P sample)
- Not in sample block on Census Day
  - Erroneous census enumerations (E sample)
  - People on P-sample rosters but not residing in sample block on Census Day
- Unresolved status for sample block
  - E sample unresolved
  - P sample unresolved

Matching AR to 'gold standard'



Occupied HUs with no census, CCM or AR records

Design scenarios for completing NRFU with AR data



United States Census Bureau

Estimate for NRFU from Scenario *i*

P_i = weighted number of people enumerated in NRFU visits *0* though *i*

+

weighted number of AR records in nonresponding HUs after *i* attempts

In contrast, 'gold standard'

- Includes only people CCM found in sample blocks
- Excludes records for people <u>not</u> in sample blocks
- Does not have records on AR list only

Measure of DQ for Scenario *i*

$D(i)$ = 'gold standard' - P_i

However, *D(i)* has to be viewed in context

- Status of the AR records found by matching to the 'gold standard'
  - Large number of AR records 'not in sample block' may offset deficiencies in correct, unresolved, & nonmatching AR records
- Measurement error in 'gold standard'

Measurement errors that may affect *D(i)* for Scenario *i*

- Coverage of CCM enhanced list
  - HUs not interviewed in census or CCM but compensated for by missing data methods
  - Persons that both the census & the CCM miss
- Nonsampling errors in CCM
  - Systematic reporting errors about moves close to Census Day may affect CCM status
    - Largest source of error in past evaluations

Errors in move dates around Census Day create errors in ‘gold standard’

Out-mover

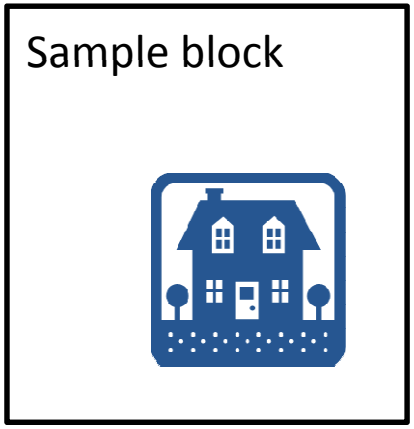
moves out of sample block after Census Day but before CCM Interview Day.
But, if reported move date is before Census Day, CCM has person not in sample block when person was.

Stable Resident

lives in sample block on Census Day & CCM Interview Day

In-mover

moves into sample block after Census Day.
But, if reported move date is before Census Day, then CCM has person in sample block when person was not.



Estimating cost *C(i)* for Scenario *i* Considerations

- Set up of field offices across country to manage a temporary workforce for NRFU
  - Or, is there a design that requires fewer temporary offices?
- Recruiting, testing, hiring, training temporary workforce
  - Fewer attempts may require smaller workforce but infrastructure for staffing may not be much different
- Questionnaire processing – even if electronic
- Creation of AR list from merging several sources of AR records

Concluding remarks

- Provided framework for building empirical cost models to assess alternative methods for a statistical program
  - Example illustrates empirical cost model: *(D(i), C(i))* for each Scenario *i* for completing NRFU with AR data
- Since census numbers used to distribute “fixed pie” resources, need measure of DQ for distributions across groups & geographic areas
- Scenarios highlight how error in ‘gold standard’ may affect assessment of DQ

Contact

mary.h.mulry@census.gov

bspencer@northwestern.edu