Merging Administrative Records in the Absence of a Register: Data Quality Concerns and Outcomes of an Experiment in Administrative Records Use

Dean H. Judson

Planning, Research and Evaluation Division

Basic Conclusions and Recommendations

- An administrative records census is not currently feasible, but might be for 2020
- Numerous applications of administrative records research for 2010:
  - Nonresponse Follow-up (NRFU) substitution
  - Imputation methods improvement
  - Master Address File (MAF) targeting
  - Census unduplication confirmation
  - Large scale data processing and record linkage improvements
  - Social Security Number (SSN) verification and search
  - Population estimation, survey improvement
- Four major improvements needed for an Administrative Records Experiment in 2010 (AREX 2010):
  - Race and Hispanic origin improvements (underway)
  - Timeliness of AR data (harder, but doable)
  - Greater geographic representativeness
  - Experimental variation on key design dimensions

What Was the Purpose of AREX 2000?

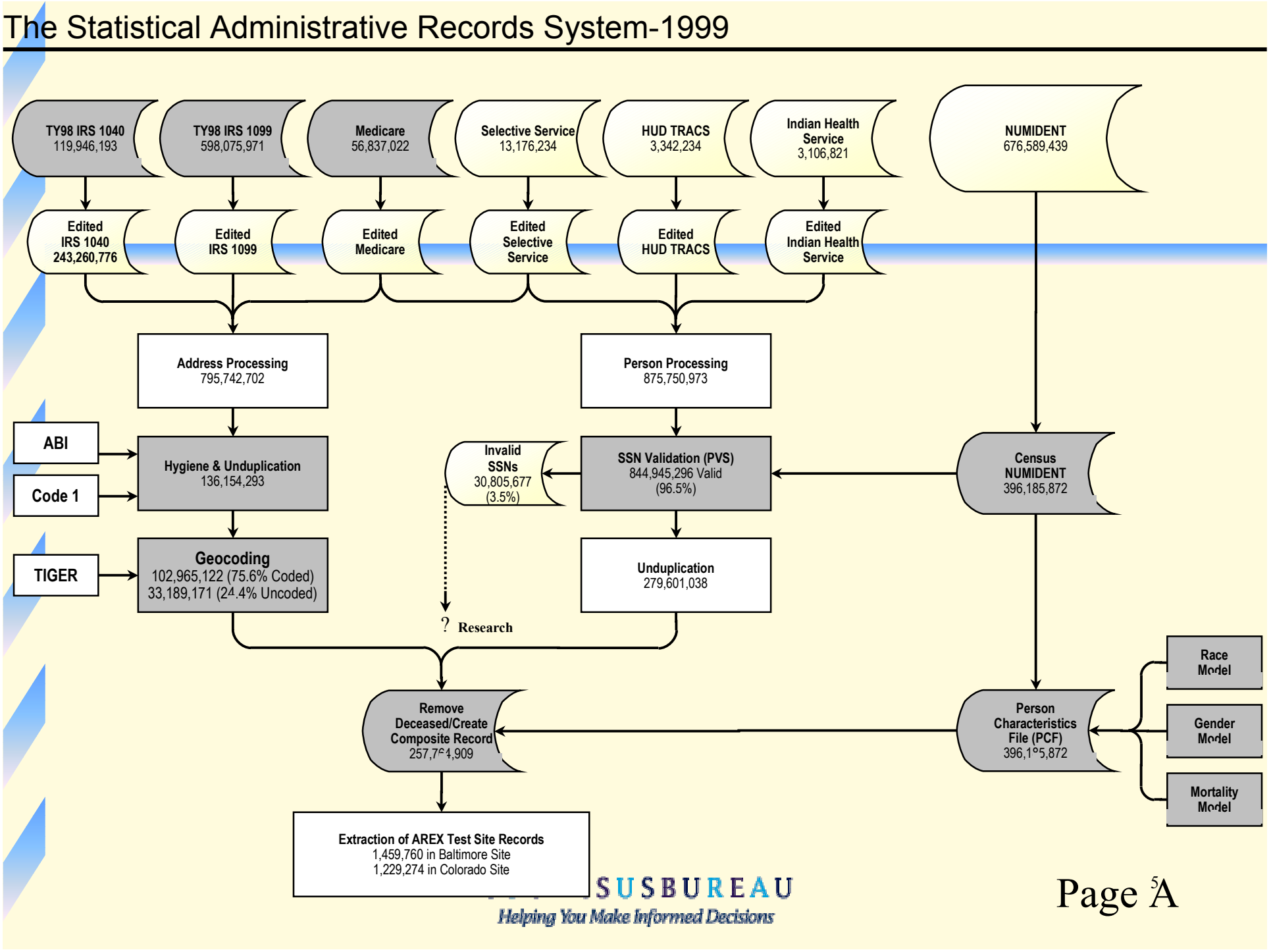
- Test the feasibility of an administrative records census
  - two counties in Maryland
    - 1.4M persons in 558,000 households
  - three counties in Colorado
    - 1.2M persons in 459,000 households
- Test two methods for conducting an administrative records census
  - top-down method
  - bottom-up method (match to address list, additional operations)

The Foundation of AREX: The Statistical Administrative Records System (StARS)

- Prototype based on 1998 vintage files
- Census-like structure and content
- Final database comparable Census 2000

  (Flowchart page A)

The Statistical Administrative Records System-1999



SUSBUREAU

Helping You Make Informed Decisions

Administrative Source Files (StARS)

- Internal Revenue Service tax files
- Medicare enrollment database
- Public housing assistance file
- Selective Service registration file
- Indian Health Service file
- Social Security Number master file (NUMIDENT--lookup file)

(Refer to Tables, pages B,C)

Creating Final StARS Database

- Select best address and demographics based on
  - geocodability
  - currency
  - quality
- Impute missing demographics
- Flag records for deceased people
- Final database is like the census

Address Processing Results (StARS)

- Almost 800 million addresses at start
- About 6 percent identified as potential businesses
- 136 million address records after unduplication
- About 75 percent geocoded
  - 85 percent geocoding rate for city-style addresses

Person Processing Results (StARS)

- 875 million records at start
- 845 million have valid SSN record (96.5%)
- 280 million after unduplication by SSN
- 261 million after removal of known deceased
- 257 million after removal of known deceased and persons residing in outlying territories

(vs. Census 2000 population of 281 million)

Additional Operations of AREX 2000

- Clerical geocoding
- Request for physical address (for P.O. Boxes, Etc.)
- Match to Decennial Master Address File
- Field address verification

(refer to AREX flowchart, page D)

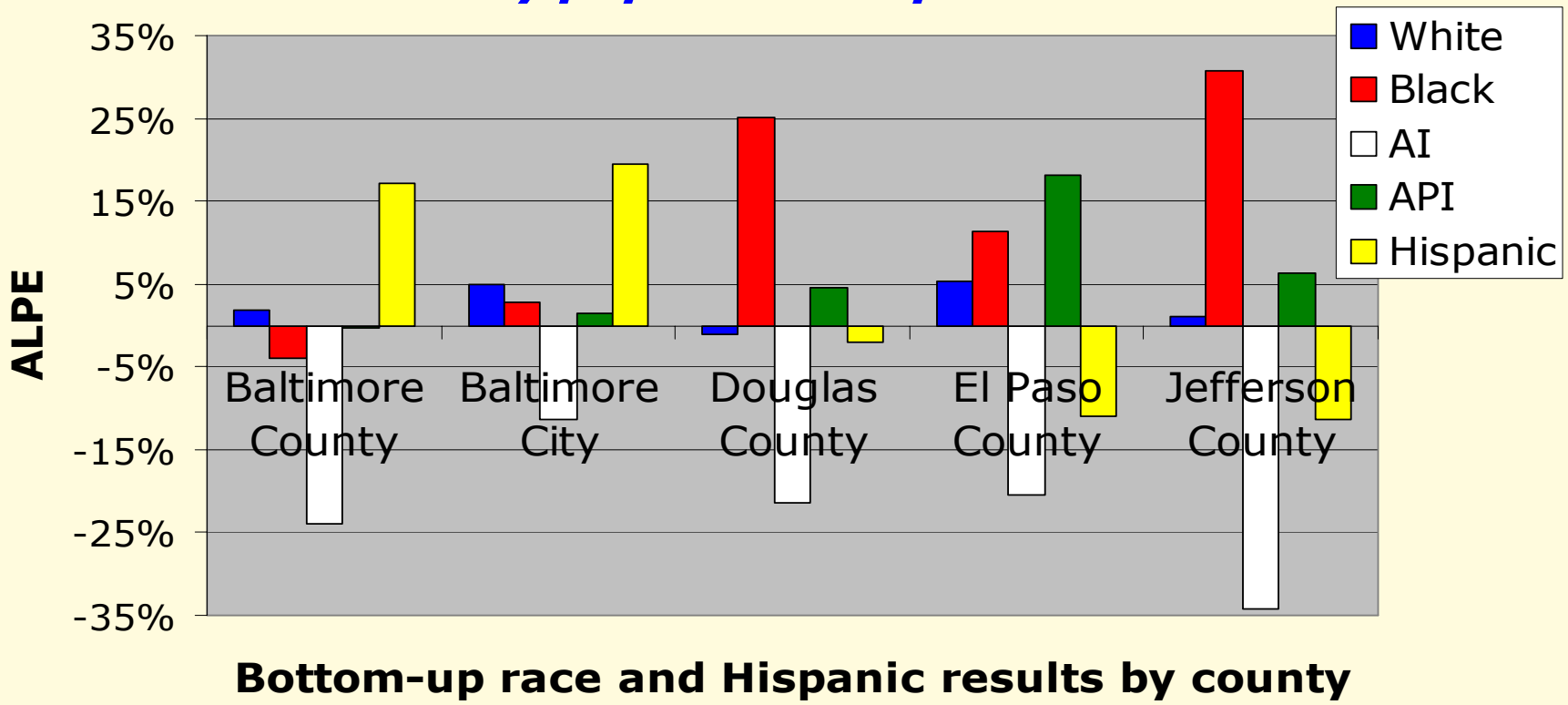
County Results: Total Population

	Census 2000	Top-Down		Bottom-Up	
		A-C Diff	%	A-C Diff	%
Total	2,558,212	-212,725	92	-25,763	99
Baltimore County	736,652	-40,469	95	-8,447	99
Baltimore City	625,401	-54,753	91	11,328	102
Douglas County	175,300	-27,030	85	-5,660	97
El Paso County	501,533	-44,642	91	-7,280	99
Jefferson County	519,326	-45,831	91	-15,704	97

Note: A=AREX count; C=Census count

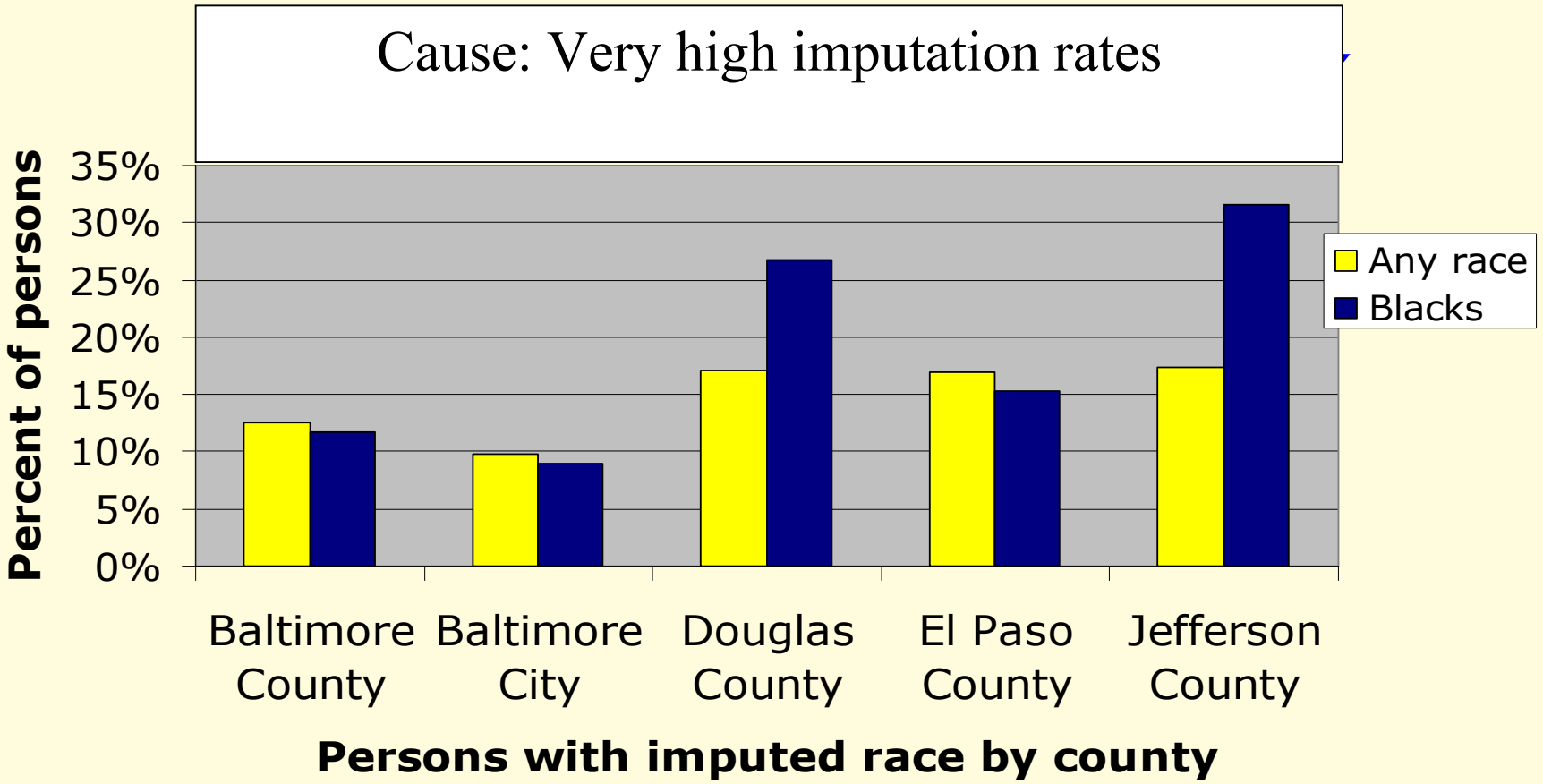
County Results: Race and Hispanic Algebraic Percent Error (A-C)/C

Size of minority population impacted ALPE results.



Bottom-up race and Hispanic results by county

County Results: AREX Race Imputation



Additional County Results

- Bottom-up performed better than top-down
- Children undercounted
- Aged overcounted
- Evidence of imputation problems

Tracts and Blocks

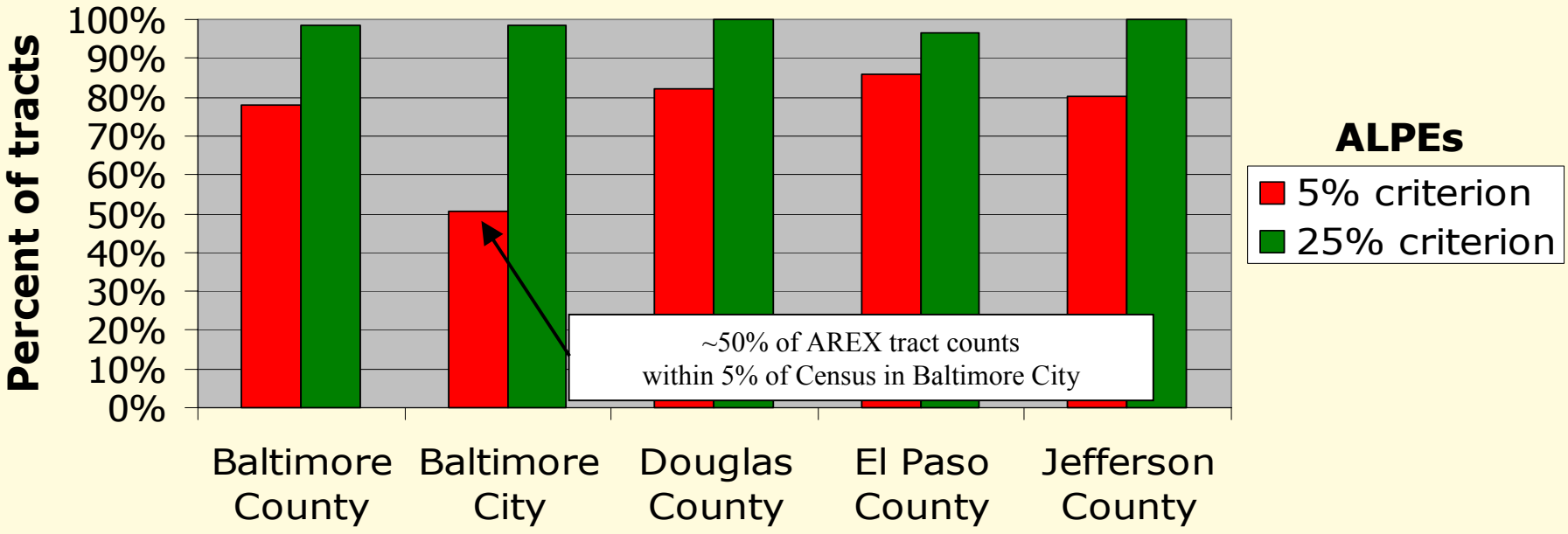
- MD site: 404 tracts, 17,041 blocks
- CO site: 283 tracts, 22,945 blocks

…

- Emphasis on ALPE { *(A-C)/C* } distributions
- 5% criterion: AREX is +/-5% of Census
- 25% criterion: AREX is +/-25% of Census

Tract results: Total Population

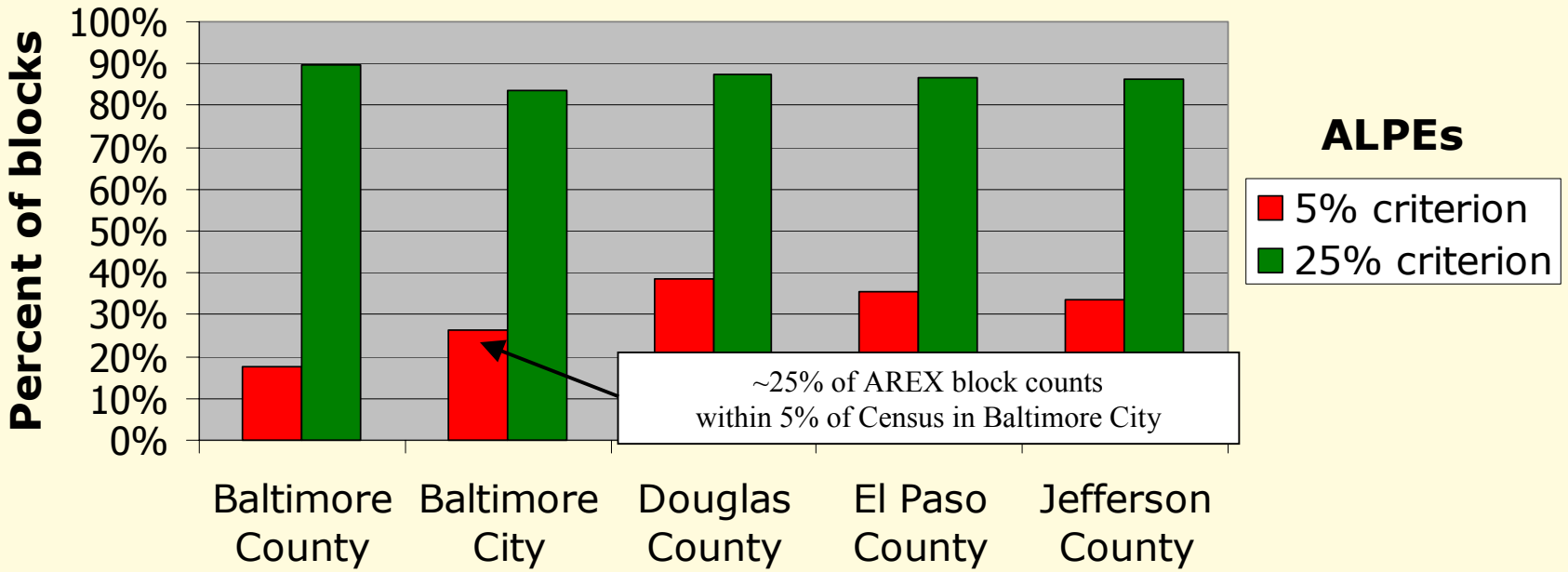
Results are more stable in Colorado;Highly variable in Baltimore County and City



Tract ALPE distributions by county

Block Results: Total Population

Results notably less stable at the block level



Block ALPE distributions by county

USCENSUSBUREAU
Helping You Make Informed Decisions

Household Evaluation Questions

- Can we use administrative records data to contribute to census operations at the household level?
- Analytic questions:
  - Do AREX addresses (computer) link to different kinds of Census addresses at varying rates?
  - Do AREX addresses that (computer) link to Census addresses contain the same number of people?
  - Do AREX addresses that link to Census addresses contain similar demographic distributions (demographically match)?
  - Can we *predict*, using non-Census information, *when* an address will demographically match?

Link Rates

- How often did AREX addresses link to Census addresses?
  - 81.4% of census addresses linked
    - ~ 5% more “imperfectly” linked
  - higher (84.0%) in occupied
  - lower (46.4%) in vacant

Link Rates and Coverage: By NRFU Status

Coverage by AREX of Census housing units, by NRFU status*

Type of Census housing unit	Total	Linked with AREX housing units
NRFU	360,914	70.9%
non-NRFU	716,450	88.4%
Occupied NRFU	289,224	76.7%
Occupied non-NRFU	715,115	88.5%
Vacant NRFU	71,690	47.6%
Vacant non-NRFU	1,335	58.7%

Link Rates and Coverage: By Imputation Status

Coverage by AREX of Census housing units, by imputation status

Type of Census housing unit	Total	Linked with AREX housing units
Imputed	24,584	62.3%
Non-imputed	1,067,876	81.9%
Imputed occupied	23,811	63.2%
Non-imputed, occupied	993,462	84.5%
Imputed vacant	773	34.7%
Non-imputed, vacant	74,414	46.5%

Number of People in Linked Addresses

- Do AREX addresses that link to Census addresses contain the same number of people?
  - Equal number: 51.1% of all linked addresses
  - Plus/minus one: 79.4% of all linked addresses

Matching Demographic Distributions in Linked Households

- How do the demographic properties of linked households compare?
- Demographic categories:
  - Sex
  - Race (4 groups, with Census multirace allocated)
  - Hispanic origin
  - Age (5 year categories and 0-17, 18-64, 65+)
  - ARSH—age, race, sex and Hispanic origin

Demographic Distributions: Overall

Comparisons between AREX and Census for demographic groups, for linked households with the same number of people only.

HH Size	Total linked, of equal size	Equal for all sex groups	Equal for all race groups	Equal for all Hisp. groups	Equal for all 5-year age groups	Equal for age groups 0-17, 18-64, 65+	Equal for all demographic groups
All sizes	445,426	91.2%	93.4%	94.8%	81.3%	93.1%	80.5%
1							85.4%
2							84.3%
3							72.2%
4							74.0%
5							69.5%
6							59.2%
7+							28.7%

Prediction

- Predicting Where An Arex Household Will Be Similar To A Census Household
  - Goal: Use *only* information available *prior* to Census MO/MB and NRFU operations to predict which addresses will match accurately
  - If we can predict ***well***, then we can use that predictive model to tell us which addresses are good candidates for NRFU substitution or imputation
  - Exploratory logistic regression model

Prediction: Simple Relationships

Address contains only persons 65 and older versus demographic match/nonmatch status.

	All AREX persons age 65 or older? No	Yes	Total
Nonmatch	513,926 66.56	33,418 28.43	547,344
Match	258,150 33.44	84,144 71.57	342,294
Total	772,076 86.79	117,562 13.21	889,638 100

Difference of proportions

Logistic Regression Model Results

- Estimated odds ratios
  - Single unit: 2.6
  - One or two persons in HH: 3.5
  - No AREX imputed race: 2.1
  - AREX one or more white: 2.1
  - All AREX 65 and older: 1.7
- Interaction effects:
  - Total effect of 65+,nonmulti,nonimputed: ***5.2***
  - Total effect of 65+,1+white, 1-2 persons: ***19.2***

Conclusions, continued

- Top-down results are not sufficient for enumeration at the block level
- Bottom-up results were better
- Tract, block results differential and predictably less accurate
- AREX covered the universe of Census HUs well
- Comparisons of household size and demographic composition were relatively promising
- AREX and Census household level characteristics were less similar for NRFU HUs
  - Open questions: Role of vacant addresses, quality of NRFU data.

Recommendations

- **Continue development now**
  - Build/train team; build acquisition relationships
- **Improve race and Hispanic origin data**
  - Underway—NUMIDENT race enhancement
  - Continue to update NUMIDENT with ACS (others?)
- **Improve timeliness of AR data**
  - Obtain IRS/other agencies' data on a ***flow*** basis
  - SSA W-2's
  - Birth/death data
  - Additional data sources

Recommendations

- **Improve geographical representativeness of AREX 2010**
- **Implement experimental variation on key processing dimensions**