Maximizing Linkage in Address Data: Spatial, Exact, and Fuzzy Matching

Timothy Champney, The MITRE Corporation
Hongxun Qin, The MITRE Corporation
Stephanie Coffey, U.S. Census Bureau

FCSM College Park, MD

Tuesday, October 22, 2024

Disclaimer: Any opinions and conclusions expressed herein are those of the author(s) and do not reflect the views of the U.S. Census Bureau. The Census Bureau has ensured appropriate access and use of confidential data and has reviewed these results for disclosure avoidance protection (Project 7527794: CBDRB-FY25-CES025-002).

Background

- High quality record linkages are critical for enriching survey data with alternative data sources
- Assigning unique identifiers at the household- or person-level enable direct linkages across datasets
- Census Bureau has an internal procedure, Person Identification Validation System (PVS), that creates these identifiers
- Linked alternative data can be used throughout the survey lifecycle, in sampling, data collection operations, and post collection processing
- Not all records in alternative data sources are successfully assigned identifiers

Background, cont.

- This work is:
  - Part of a larger portfolio of work to leverage administrative and commercial data to reduce respondent burden and increase data collection efficiency
  - Focused on address-based linkages between the American Community Survey (ACS) and third-party sourced real estate (RE) data
- Multiple methods to assign Master Address File IDs (MAFIDs)
  - Census Bureau linkage process
  - Other methods to add additional linkages
- The key is to maximize linkages between Census Bureau household survey samples and alternative data sources

MAF and MAFID

- The Master Address File (MAF) is a Census Bureau file that contains an inventory of all known living quarters in the country
- Each physical address is assigned a unique identifier, MAFID, in the Master Address File
- The MAF is updated twice each year using the previous MAF and USPS files including the Delivery Sequence Files (DSF), ZIP Move Engineering File, and Locatable Address Conversion System, etc.
- Records represent a single structure or unit within a structure and include housing units, group quarters, and nonresidential

Problem Statement

- How to improve on the basic MAFID linkage between household level Census Bureau survey data and third-party data?
  - Third-party data are introduced to enhance Census Bureau survey data
  - The third-party data have multiple files on the homes/properties of the country and one of the files contains the information of interest
  - Some third-party data also have shape files of the property boundaries that can be used to refine matching
  - The same algorithms were used to assign MAFIDs to the properties in the third-party data using the address information as in the ACS
  - The base case is to link Census Bureau survey data to one third-party file

Previous Research

- Binder et al (2022): Compared American Housing Survey (AHS) 2019 data to two commercial vendor property data sources (2019 and 2017 data respectively)
  - Assessed match rates based on MAFID with rates enhanced by fuzzy matching and matching to geographic boundary files
  - Fuzzy matching was applied to records not having a MAFID match
  - Higher agreement rates were found with owner-occupied than renter occupied units
  - Matching rates were better for single-family housing units than multi-family or other types of units. When added as a matching step, the spatial matches improved coverage by 7-9%.
- Dillon (2019): Compared American Community Survey (ACS) data to a single vendor's property data for 2014
  - Found lower linkage rates among sources for households with young children, minorities, residents of group quarters, recent movers, low-income residents, the unemployed, rural residents, and occupants with low education
- Both studies
  - Found similar MAFID linkage rates (60-67%).
  - Noted limitations in vendor data such as source originating with owners verses the occupants in surveys and that property tax records my reference an entire parcel verses a housing unit

Data Sources

- American Community Survey (ACS) 2019 unswapped, unweighted housing unit records with MAFIDs
- Commercial data vendor 2019 property records preprocessed by the Census Bureau with MAFIDs attached

Base Case: One Third-Party Data File Matched to ACS on MAFID

Census Survey Data			Third Party Data		
MAFID	Address Info	Data	MAFID	Address Info	Data

Two Approaches to Improve Linkage

- Utilize multiple third-party data files
- Apply Geo-Spatial Matching and Address Fuzzy Matching

Multiple Third-Party Files vs One

- The properties in all third-party data files are assigned MAFIDs
- But not all properties are assigned MAFIDs
- Further the MAFIDs for the same unique ID may not be the same
- Simple VOTE/RANK methods are used to give a third-party property ID a unique MAFID
- Base rate:
  (# matches in base case/# ACS sample household units)
- The improvement rate over the base case is calculated:
  (# new matches/# matches in base case)

United States® Census Bureau

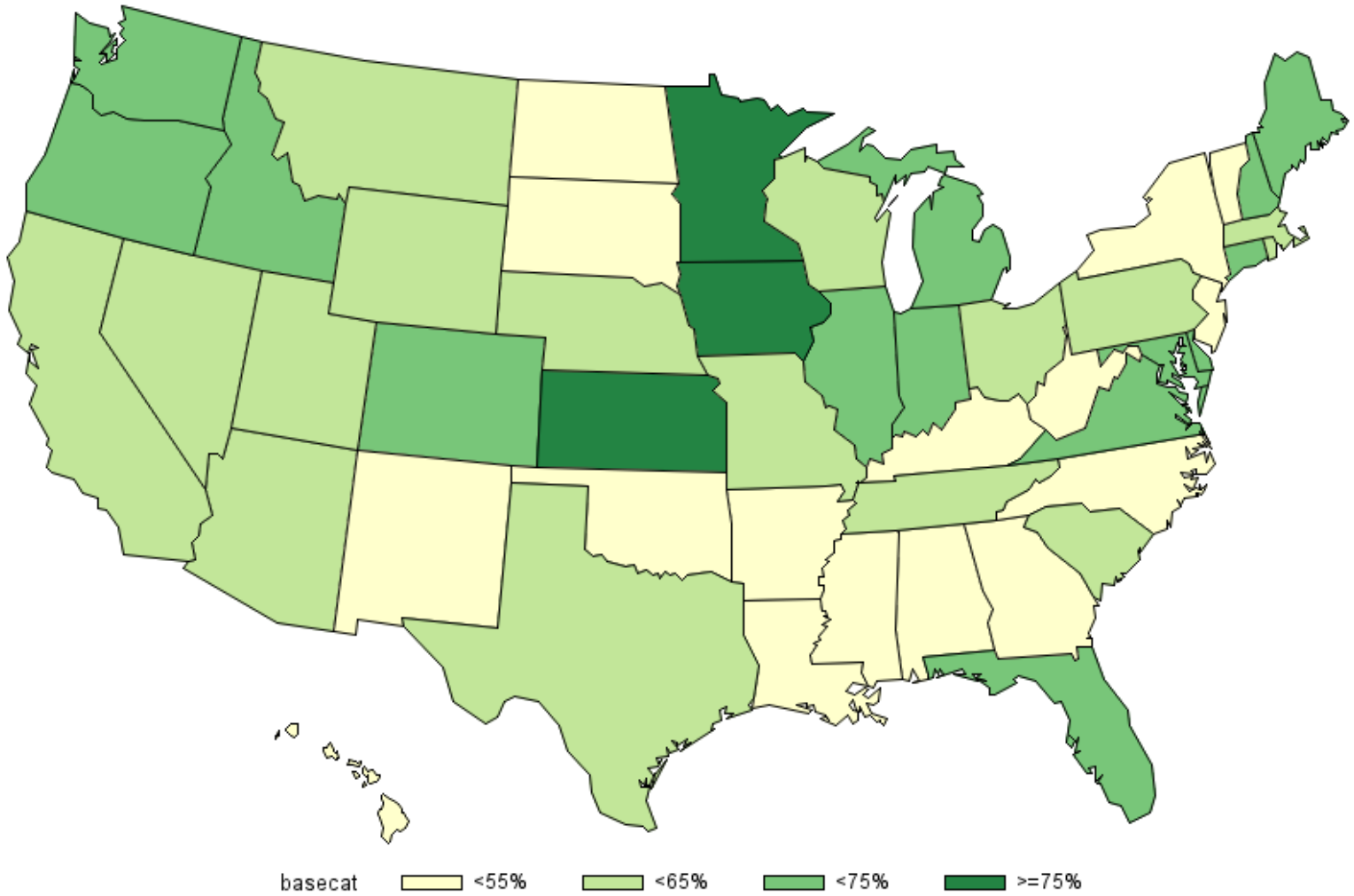
Enhancement 1: Multiple Third-Party Data Files Used to Determine Best Match

Base File	Unique ID	MAFID	Address Info	Data
Other File 1	Unique ID	MAFID		
Other File 2	Unique ID	MAFID		
Other File 3	Unique ID	MAFID		
…				

United States® Census Bureau

Unweighted Linkage Rates, Before Enhancements

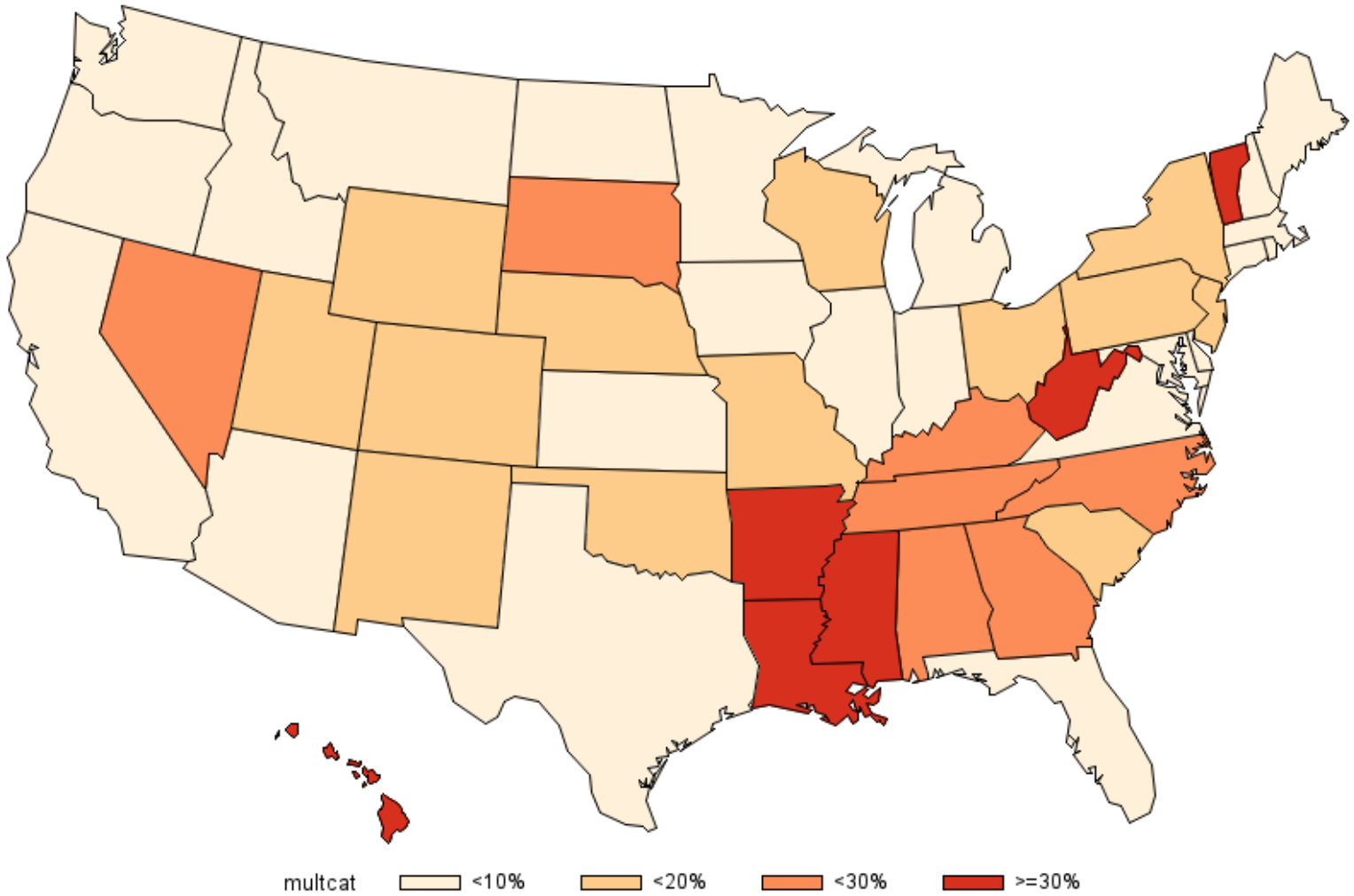
Address-Based Linkage Between 2019 ACS and Third-Party Data



United States® Census Bureau

Unweighted Linkage Improvement Rates Using Multiple Files

Address-Based Linkage Between 2019 ACS and Third-Party Data



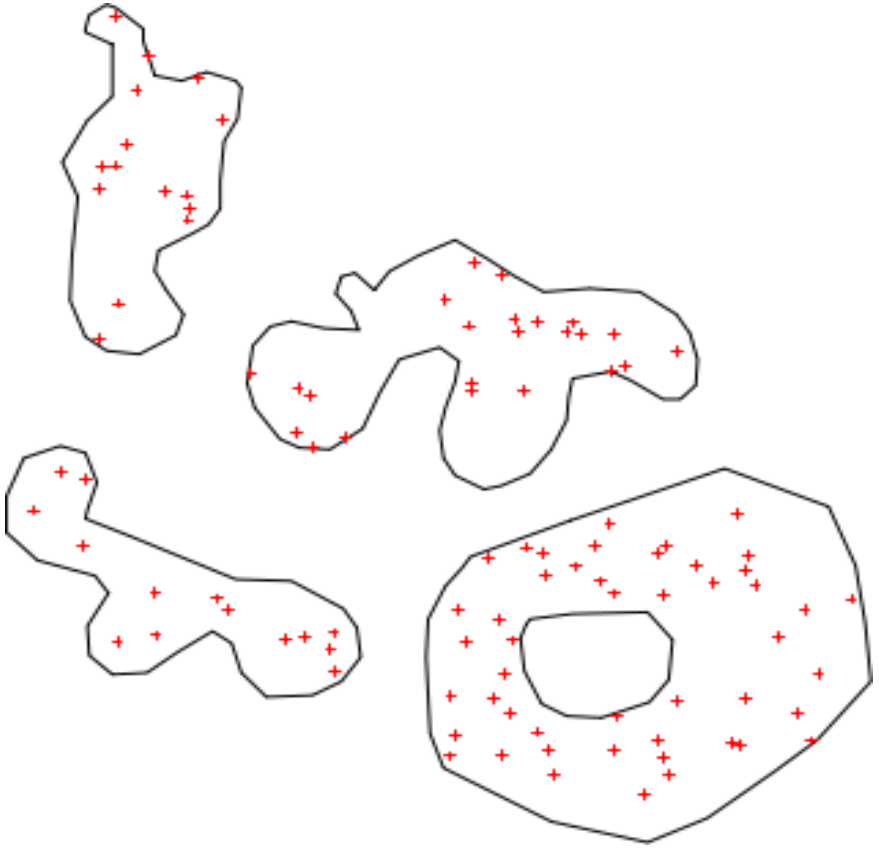
United States® Census Bureau

Enhancement 2: Geo-Spatial Matching

- Properties in Census Bureau survey data have latitude and longitude
- The survey properties (latitude/longitude) can be matched to the third-party properties (shape files) using an R Geo-Spatial package
- Since the matching is by approximation, false matches are introduced, especially for multi-unit properties such as apartments and condos
- Outputs from the Geo-Spatial matching need to be further cleaned

United States® Census Bureau

Graphic Illustration of Geo Shape File Matching Points to Polygons



United States® Census Bureau

Geospatial Matching Alone is Not Enough

	Same House Number?	
House Type	YES	NO
Multi-Unit	55.4%	44.5%
Single Family	86.0%	13.9%
Trailer	55.5%	44.4%
Other	53.4%	46.6%
No Value	66.8%	33.1%
Overall	77.8%	22.1%

United States® Census Bureau

Fuzzy Matching on Address

- For the outputs from the Geo-Spatial matching, house number, nonempty unit number, and street name are used to filter out false matches
- Naturally the house numbers should be the same
- If either side (Census Bureau or third-party data) has non-missing unit number, then the unit number should be the same
- Different distances on street name are analyzed and cutoff thresholds are used for filtering, hence the term Fuzzy Matching

United States® Census Bureau

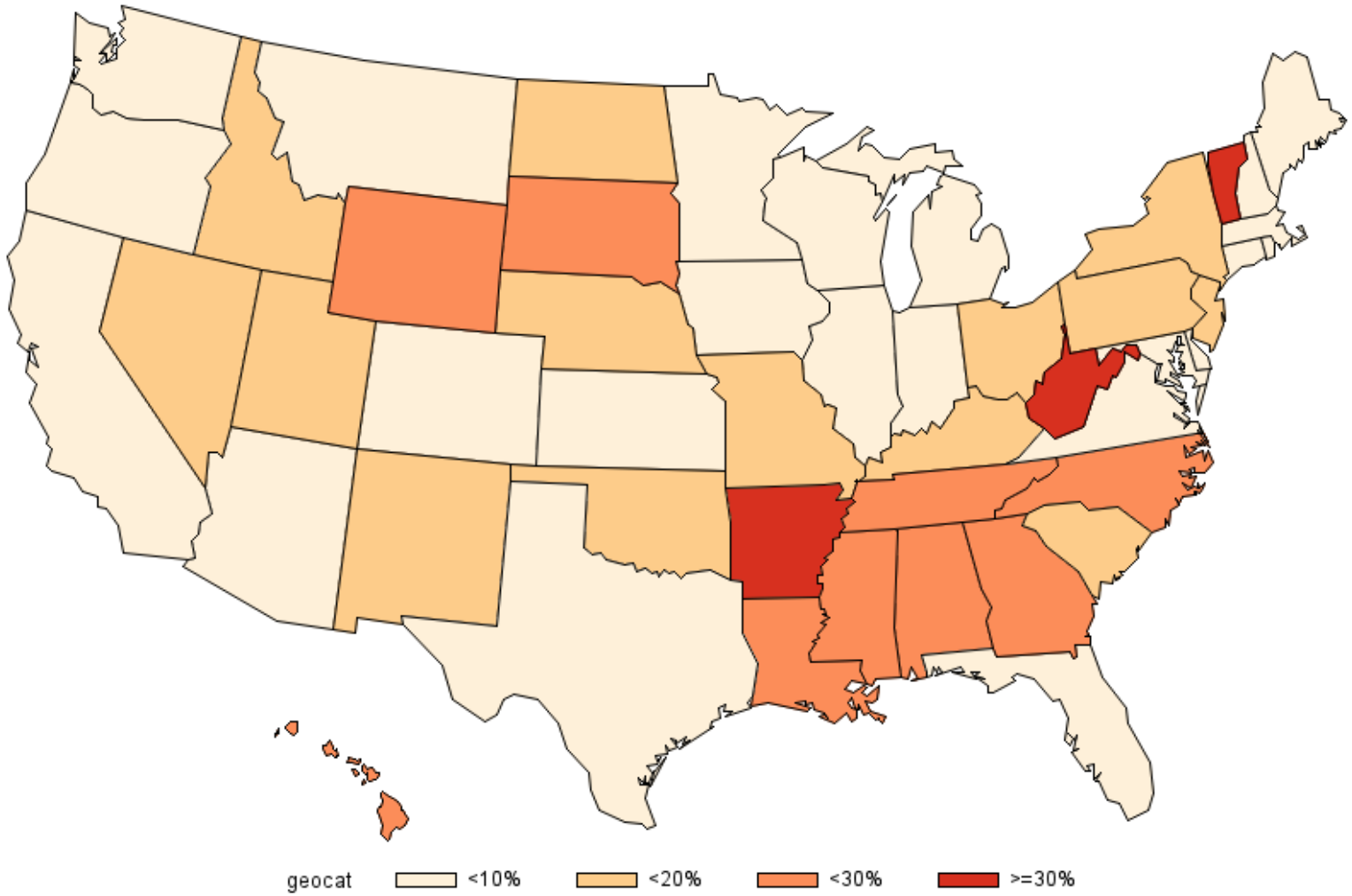
Example of Fuzzy Matching

House No.	Street	Unit No.	House No.	Street	Unit
231	Tollgate		231	Tollgate	
401	5th		401	Fifth	
12329	Allensville Ave		12329	Allensville	
888	CO 115		888	County 115	
841	Grand Valley	103	841	Grand Valley Pointe	103

Note: exemplar street addresses

Unweighted Linkage Improvement Rates Using Geospatial Linking

Address-Based Linkage Between 2019 ACS and Third-Party Data



United States® Census Bureau

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

Two Approaches to Improve Linkage

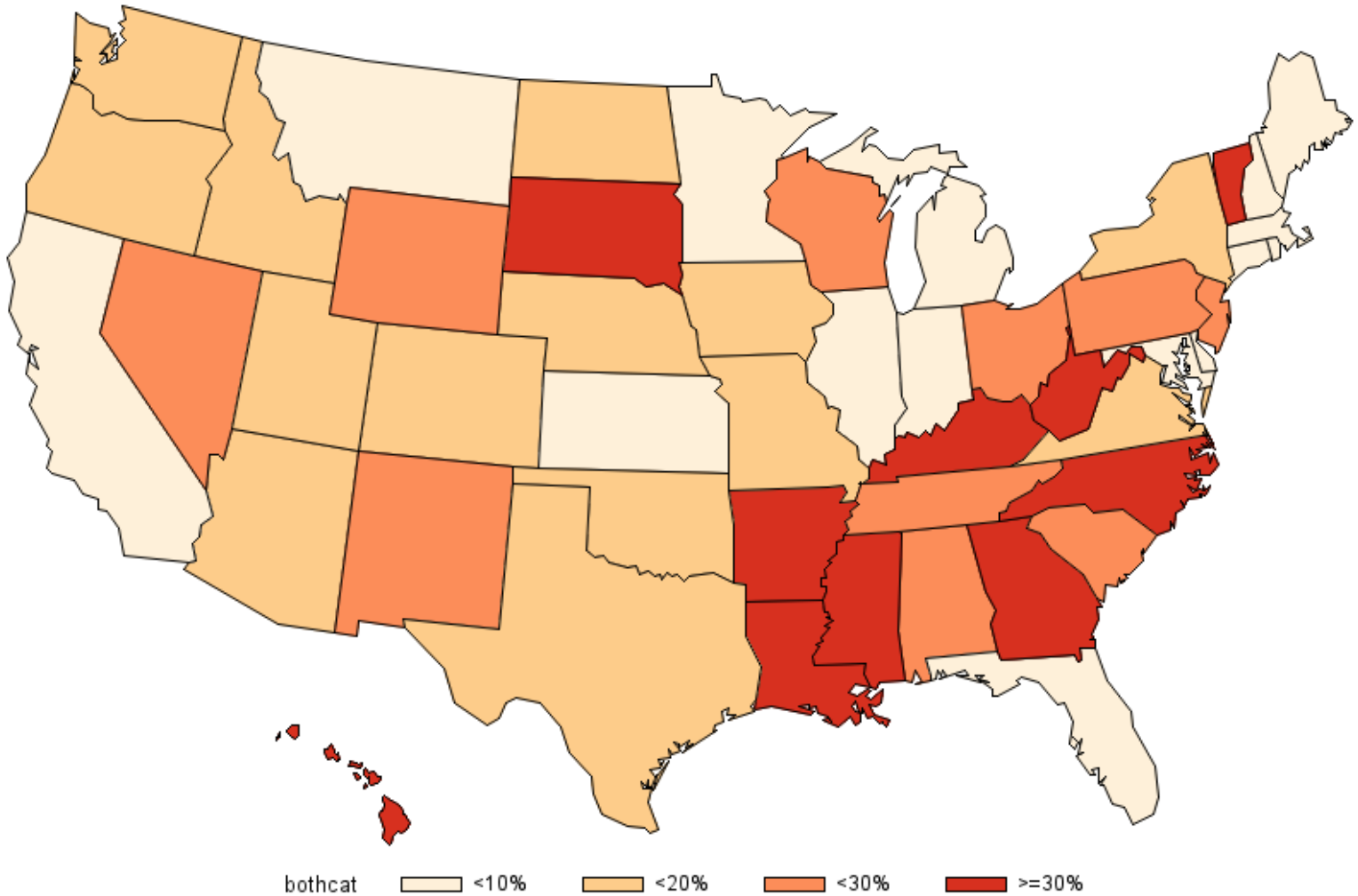
- Utilize multiple third-party data files
- Apply Geo-Spatial Matching and Address Fuzzy Matching
- Though the analysis is conducted independently, in practice, the two approaches can be applied sequentially
- The two approaches have heavy overlaps
  - Hence, when applied sequentially, the improvement is not the sum

Base Rates and Improvement Rates

	Base Rate	Improvement Rate			Final Rate
House Type	Base File	Multiple File (Alone)	Geo-Spatial+ Fuzzy Match(Alone)	Both with Cleaning	With Combined Improvements
Multi-Unit	17.1%	11.9%	6.0%	14.4%	19.6%
Single Family	78.1%	8.0%	9.7%	12.7%	88.1%
Trailer	42.5%	17.5%	18.9%	30.3%	55.3%
Other	19.3%	6.3%	35.2%	32.0%	25.5%
No Value	42.7%	5.6%	7.5%	14.9%	49.1%
Overall	61.6%	8.5%	9.8%	13.7%	70.0%

Unweighted Linkage Improvement Rates Using Both Methods

Address-Based Linkage Between 2019 ACS and Third-Party Data

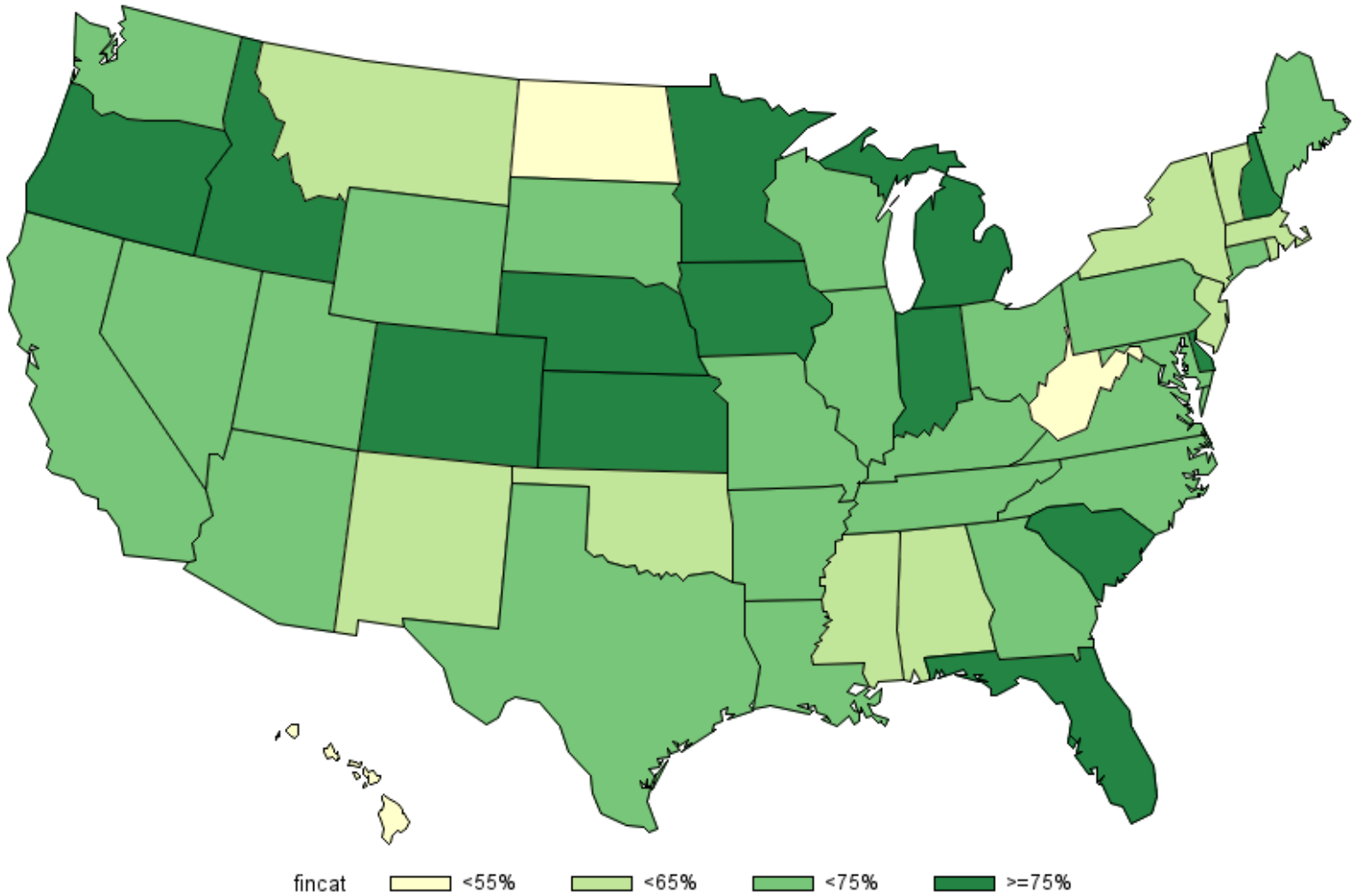


United States® Census Bureau

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

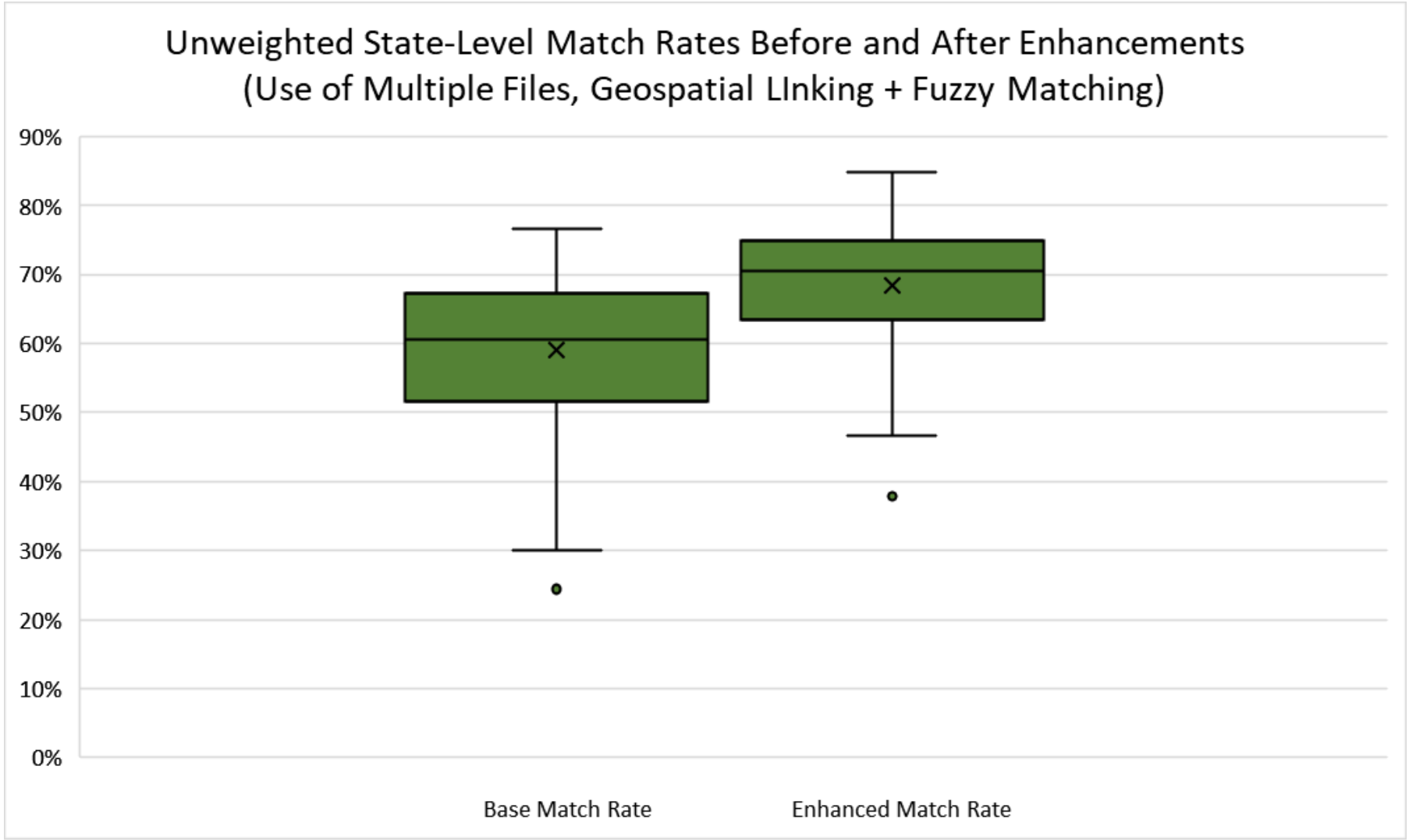
Unweighted Linkage Rates, After Enhancements

Address-Based Linkage Between 2019 ACS and Third-Party Data



United States® Census Bureau

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®



United States® Census Bureau

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

Findings and Conclusions

- Each of the two approaches improve the matching rates between Census Bureau household survey and third-party data
- Due to the different natures of house structures, the improvements differ for different types of properties
- Improvements for Multi-Units and Trailers are especially significant as the Census Bureau household survey data have low coverage
- Other data sources should be considered in the future

References

Binder, A.J., Molfino, E., & Voorheis, J. (2022) Comparing the 2019 American Housing Survey to Contemporary Sources of Property Tax Records: Implications for Survey Efficiency and Quality https://www2.census.gov/ces/wp/2022/CES-WP-22-22.pdf

Brummet, Q. (2014) Comparison of Survey, Federal, and Commercial Address Data Quality (census.gov) https://www.census.gov/content/dam/Census/library/working-papers/2014/adrm/carra-wp-2014-06.pdf

Clark, S.L. & Sawyer, R.C. (2018) Housing Administrative Records Simulation (census.gov) https://www.census.gov/content/dam/Census/library/working-papers/2018/acs/2018_Clark_01.pdf

Dillon (2019) Preliminary Research for Replacing or Supplementing the Acreage, Number of Rooms and Bedrooms, Tenure, Property Value, & Real Estate Taxes Questions on the American Community Survey with Administrative Records (census.gov) https://www.census.gov/content/dam/Census/library/working-papers/2019/acs/2019_Dillon_01.pdf

Layne, M., Wagner, D., & Rothhaas, C. (2014) Estimating Record Linkage False Match Rate for the PVS (census.gov) https://www.census.gov/library/working-papers/2014/adrm/carra-wp-2014-02.html

Pebesma, E. & Bivand, R.S. (2005) Classes and Methods for Spatial Data: the sp Package https://cran.r-project.org/web/packages/sp/vignettes/intro_sp.pdf

Pebesma, E. & Bivand, R. (2023). Spatial Data Science: With Applications in R (1st ed.). Chapman and Hall/CRC, Boca Raton. https://doi.org/10.1201/9780429459016

SAS (2021). SAS/GIS® 9.4: Spatial Data and Procedure Guide. https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/apdatgis/titlepage.htm.

Seeskin, Z.H. (2016) Evaluating the Use of Commercial Data to Improve Survey Estimates of Property Taxes https://www.census.gov/content/dam/Census/library/working-papers/2016/adrm/carra-wp-2016-06.pdf

Wagner, D. & Layne, M. (2014) The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software (census.gov) https://www.census.gov/content/dam/Census/library/working-papers/2014/adrm/carra-wp-2014-01.pdf

United States® Census Bureau

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®