Journal of Data Science 24 (3), 600–615
July 2026

DOI: 10.6339/26-JDS1236
Data Science in Action

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

Timothy F. Champney[1,*] and Hongxun Qin[2]

[1]*Formerly The MITRE Corporation, Human Systems Integration and Analysis, Washington, DC 20008, United States*

[2]*The MITRE Corporation, Federal Statistics and Data Science, McLean, VA 22102, United States*

Abstract

High quality record linkages are critical for enriching survey data with alternative data sources. To enhance the American Community Survey (ACS) (US Census Bureau, 2025), we evaluated several options to match commercially available property data to housing unit records in the ACS and the Census Master Address File (MAF) (US Census Bureau, 2022) data, with the ultimate goal of supplementing data collected in the survey with the commercial property data. Techniques to match address data come in two flavors: spatial matching and address matching. Spatial matching is done by overlaying commercial boundary shape files on the lat-long coordinates on the MAF to associate Census housing unit records to commercial property parcel records. This method is useful because it does not require matching of text fields, but performs poorly when parcels include many housing units (e.g., large apartment buildings). Address matching, or entity resolution at the address level, links records across the two sources based on the content of various address fields, and offers the possibility of disambiguating multiple matches and matching objects that cannot be successfully assigned a unique match through spatial matching. Several approaches are available for address matching, including rule-based, deterministic, fuzzy, and probabilistic methods. In our article we summarize literature comparing various combinations of spatial and address matching techniques to illustrate the trade-off between linkage rates and linkage quality. We also consider hybrid solutions that leverage spatial matching as well as several types of address matching to maximize high quality linkages for our research and discuss future directions such as incorporating probabilistic matching as an added step.

Keywords *ACS; address list; Census; deterministic matching; fuzzy matching; MAF; probabilistic matching; spatial matching*

1 Introduction

Augmenting survey data with data from address-level administrative data requires accurate matching of addresses between data sources (Comber and Arribas-Bel, 2019). A variety of commercially available administrative property datasets have the potential for matching with the U.S. Census American Community Survey (ACS) household level records and may contain data elements that respondents to the survey are unable or refuse to report. These data elements may include assessed acreage, property values, property taxes, tenure, home owner status, numbers

☆Approved for Public Release, Distribution Unlimited. Public Release Case Number 25-2360.

*Corresponding author. Email: tfchampney@earthlink.net.

Received September 8, 2025; Accepted May 29, 2026

of rooms and bedrooms, kitchen and bathroom facilities, and other amenities.

The Census Bureau has sought to match this type of administrative data with ACS sample records over the past ten years with mixed success rates. A routine matching process has been put in place to attach matching identifiers to commercial property records when they are entered into the Census data system (Raim, 2016). This standard process has resulted in about a 60% match rate. When ingesting address-based administrative records, the Census Bureau matches them to the Census Master Address File (MAF), which contains MAF identifying keys (IDs) known as MAFIDs. The MAFIDs are attached to the ingested records based on the match. The MAF consists of an inventory of virtually all address locations in the United States, including residential living quarters as well as other types of properties. Records represent a single structure or unit within a structure and include single and multiple housing unit dwellings, group quarters, and nonresidential structures, such as properties solely used for commercial purposes. The MAF is updated twice each year using the previous version of the MAF and United States Postal Service (USPS) files including the Delivery Sequence Files (DSF), ZIP Move Engineering File, and the Locatable Address Conversion System.

We address how to improve the linkage of Census MAFID household-level property records and commercially available residential real estate records. Approaches we consider include the basic production process currently used by the Census Bureau, using property boundary files (shape files) in conjunction with GPS coordinates, incorporating multiple types of commercial files with address fields that vary in quality, and fuzzy matching techniques that could be leveraged to improve the match rates. We show how one might combine these techniques in a hybrid process to improve match rates. We also consider the use of probabilistic model-based matching that leverages machine learning and artificial intelligence and the benefits and trade-offs of deterministic versus probabilistic approaches.

2 Core Concepts and Terminology

2.1 Matching Techniques

2.1.1 Rule-Based or Deterministic Matching

Deterministic matching, often referred to as rule-based or exact matching, is a widely used approach for linking address records across datasets. This method determines record matches based on the exact agreement of identifiers or specific rules, rather than relying on probability scores. Deterministic algorithms use explicit rules to determine matches. These rules may include requiring exact, character-for-character matches on all or selected fields (such as street, city, and postal code) or matching on a unique identifier if available (Dusetzina et al., 2014) (Binette and Steorts, 2022). For valid and reliable matching it is important to clean, pre-process, and regularize address fields (Placekey, 2024). The typical workflow in deterministic matching includes the following steps:

- Standardization: Addresses are preprocessed to standardize formatting (e.g., abbreviations, capitalization, field order).
- Blocking: Candidate pairs are selected based on easy-to-match fields (e.g., ZIP code, house number), reducing the comparison space and computation requirements.
- Rule Application: Matching rules are applied in ranked order—most strict first (exact string match), with subsequent rules allowing for more variation such as swapped fields or recognized abbreviations.

- Handling Variations: Deterministic methods may incorporate algorithms like Levenshtein distance, handling minor typos, plurals, or common misspellings.
- Final Classification: Pairs are classified as matches, non-matches, or possible matches based on which rule, if any, they satisfy.

Rule-based matching of address records typically requires cleaning, parsing, aligning, and standardizing address fields followed by determining matches between corresponding elements in two or more address datasets. Due to differences in spelling and alternative expressions of the same address, fuzzy matching is commonly used to determine near-matches of similar address strings (Placekey, 2024).

Fuzzy matching uses algorithms such as Levenshtein distance (edit distance), tokenization, and similarity scoring. A match threshold determines how similar two addresses must be to be considered a match—higher thresholds mean stricter matches, while lower thresholds increase recall but may admit more false positives. This flexibility makes fuzzy matching a powerful tool for address data cleansing, deduplication, and integration across multiple data sources. Fuzzy matching enhances the match rate by enabling address systems to intelligently detect, correct, and reconcile variations, typos, and inconsistencies that deterministic matching overlooks, resulting in cleaner, more complete, and more accurate address datasets (Sloan and Hoicowitz, 2016).

Alternatively geospatial matching may be applied to the datasets. Address records may be geocoded by assigning latitude and longitude coordinates to each address record and determining which pairs of records in the two datasets are in the closest proximity by geographic distance. Real estate records may also include boundary or shape files defining the perimeter of each property. When associated shape files are provided with one of two datasets to be matched, algorithms are available to identify which records have geographic coordinates within each boundary polygon in the shape file (Lee, 2025).

Geospatial methods, such as geocoding and spatial analysis, offer powerful tools for improving the match rate when linking address records—especially in large or messy datasets. These methods integrate location-based information to validate, enrich, and cross-check textual address data. In geospatial matching, textual addresses are translated into geographic coordinates (latitudes and longitudes) and back again, allowing direct comparison of physical locations. If two address records geocode to the same point or nearby points, this provides strong evidence of a match. Addresses that are not a textual match but lie within a defined spatial buffer (e.g., within 10 meters) can be flagged as probable matches, capturing true pairs that deterministic or fuzzy text matching would miss. Geospatial matching incorporates spatial datasets such as parcel boundaries, property polygons (i.e., shape files), or street centerlines, helping resolve ambiguous or partial addresses by matching them to authoritative spatial features (Lee, 2025; Tyris et al., 2024; Goldberg, Wilson, and Knoblock, 2007).

Automated geocoding and spatial validation boost match rates, with studies showing validated geocoding rates exceeding 85% and very high precision for rooftop-level matches, where rooftop-level refers to an exact match in geolocation down to the center of the rooftop (Cortes et al., 2021).

In order to overcome some of the limitations of fuzzy matching and geospatial matching, the two methods may be combined in sequence to determine a best match. Geospatial matching as a second step provides a fallback for otherwise unmatched textual addresses (Clark et al., 2023; Champney et al., 2024).

The strengths of deterministic matching include:

- High precision when data quality is high and identifiers are reliable, leading to fewer false positives.

- Transparency, because the matching process follows explicit, understandable rules and is therefore interpretable and reproducible.
- Speed, as it is typically faster than probabilistic approaches and suitable for large datasets with well-structured address information.

Limitations include:

- High sensitivity to inconsistencies, missing data, or typographical errors, often resulting in lower recall (more false negatives).
- Limited ability to handle variations in address format, abbreviations, misspellings, or duplicated properties unless complex rule sets are explicitly created.
- Reduced effectiveness for inconsistent, noisy, or incomplete data, where deterministic rules can fail to link correct records without manual intervention or extensive rule customization.

2.2 Probabilistic Matching

Probabilistic matching consists of estimating the probability of a match between each pair of records in two datasets. Probabilistic matching often employs the Fellegi-Sunter framework or probability function (Binette and Steorts, 2022). Machine learning algorithms applied to classify address pairs as matches and nonmatches can be used in place of traditional probability methods (Comber and Arribas-Bel, 2019). Probabilistic matching offers several key advantages over deterministic (rule-based) approaches, especially when dealing with large, messy, or incomplete datasets (Comber and Arribas-Bel, 2019). Probabilistic matching does not require exact field agreement, making it robust in the face of typographical errors, abbreviations, inconsistent formatting, or missing fields. It can make informed linkages based on the available information and assign a confidence score to each possible match. Probabilistic matching may consider multiple data elements simultaneously—such as name, address, date of birth—assigning weights to each field based on its relative importance or reliability. This holistic approach increases both match sensitivity and specificity. By computing a probability score and adjusting thresholds, probabilistic matching balances the risk of both false matches and missed matches. This is especially helpful in environments where perfect data quality cannot be expected. Automated scoring reduces the chance of incorrectly merging non-matching records or missing true matches compared to deterministic, fixed-rule approaches. Probabilistic models may scale efficiently to millions of records and can integrate data from multiple distributed sources with varying quality, structure, and completeness. However, probabilistic match requires applying a probability function to $(N \times (N-1))/2$ pairs of records to identify the most likely match. To reduce the number of comparisons, the data may be blocked into groups with similar address locations such as by zip code (Comber and Arribas-Bel, 2019). Studies show probabilistic linkage can reach very high sensitivity and specificity (e.g., over 99%) in real-world public health applications, even when unique identifiers are unavailable or inconsistent (Aldridge et al., 2015). Probabilistic matching and machine learning methods may be combined in sequential steps with text-based and geospatial matching, further reducing false matches (Zhang et al., 2024).

In summary, probabilistic matching may excel when data are messy, incomplete, or distributed across many sources. It increases match rates, reduces error, and adapts to a broader range of real-world data challenges than deterministic (exact) matching (Prindle et al., 2023).

2.2.1 Advantages and Disadvantages of Matching Alternatives

Address matching of units and multi-unit dwellings such as apartment buildings, condominiums, trailer courts, and multi-family homes introduces complexities that can lead to inaccurate or

incomplete record linkage or duplication. Geocoordinates and boundary shape files may be shared by all units in the same structure (Clark et al., 2023).

Deterministic matching of address records remains a preferred method in scenarios where high-quality, well-structured data are available, offering transparency and high precision. However, for complex and noisy datasets, a combination with probabilistic or machine learning methods may yield better results (Zhu et al., 2015). Other research has explored more flexible and robust rules, particularly for public sector and healthcare applications (Hagger-Johnson et al., 2017).

3 Census Bureau Work on Address Matching

Over the years, the US Census Bureau has taken a mulifaceted approach to improve matching address records. Much of the work has dealt with the address matching problem in the MAF which contains unit-level address records covering the universe of all properties in the United States, including both residential and commercial properties. Traditionally the Census Bureau has relied on partnerships with other government entities such as the USPS, Internal Revenue Service (IRS), and state and local governments to keep the MAF records up to date, e.g., the Local Update of Census Addresses (LUCA) (Moreno, 2025). Matching work has also focused on improving estimates in the ACS by matching household records in the survey to property data aggregated from county tax records and other sources by private data vendors such as CoreLogic (CL) and Black Knight(BK) (Brummet, 2014; Clark et al., 2023).

3.1 The Master Address File

The Census Bureau, Center for Administrative Records Research and Applications (CARRA) utilizes a blend of probabilistic and deterministic algorithms to match ingested address records such as CL and BK to the MAF. An index referred to as the MAFID is attached to each successfully matched ingested record. In the process the ingested records are cleaned and parsed to correct and standardize spelling, update ZIP codes, and extract address data elements such as house number, unit number, street name, prefixes and suffixes of street names or neighborhoods (e.g., rural route, or NW, ST, Ave, etc.), city names, and states. The process iterates through probabilistic and deterministic steps including direct and fuzzy matching (Brummet, 2015). The Census Bureau updates the MAF semi-annually with administrative data from the USPS Delivery Point Address List (Raim, 2016).

3.2 Key Research and Evaluation Studies

Brummet (2014) evaluated the quality of survey, federal, and commercial address data by measuring their match rates to the MAF. The analysis covered three major data sources from 2009–2011: the American Housing Survey (AHS), CL commercial property data derived primarily from county property tax records, and administrative records from the U.S. Department of Housing and Urban Development (HUD). Records from each dataset (spanning the years 2009–2011) are cleaned, standardized, and then matched to the MAF. The study used probabilistic matching routines, emphasizing the importance of address completeness and uniformity (i.e., house number, street name, suffix, unit number, ZIP code). Brummet (2014) used a two-pass approach: one for standard addresses and one for special/rural route addresses. Brummet

(2014) found AHS and HUD records show very high match rates to the MAF (85–92%), highlighting their strong address quality and completeness. In contrast, CL commercial data showed much lower match rates to the MAF (63%), reflecting more missing data and challenges matching multi-unit buildings and complex property types. Urban addresses, single-family homes, and owner-occupied properties are more reliably matched across all datasets. Multi-unit dwellings and rural addresses, or records with missing/ambiguous components (such as unit numbers), presented the greatest matching difficulties, especially for commercial sources.

Brummet (2015) evaluated various methods for linking CL data to AHS records. The study looked at ways of improving the match rate of CL data with the MAF and AHS over and above the rates reported in 2014. They found that lowering the match score threshold for fuzzy matches yielded more matches but quickly increased false matches. For instance, relaxing the street name or house number requirements significantly boosted match rates only by introducing questionable or clearly incorrect matches. Component-level penalties (for non-matches in minor fields, such as directional prefixes) have little effect due to the infrequency of these fields in real data. Blocking on only the first digit of the house number can produce moderate increases in match rates but again at the cost of more false positives. The most successful matches hinged on agreement in house number and street name. Relaxing match criteria for these key fields made substantial improvements in match rate possible, however, relaxing the match criteria simultaneously raised the risk of incorrect matches. Thus relaxing various criteria yielded match rates between 78% and 85%, however false matches also increased.

Seeskin (2016) investigated the potential for using CL 2008-2010 commercial property tax data to improve the estimation of property tax amounts in the 2010 ACS, with a focus on single-family, owner-occupied homes. The analysis addressed three main goals: reducing respondent burden by possibly replacing survey questions with administrative records, understanding ACS response error, and improving nonresponse adjustment using commercial data. Seeskin found that the coverage and data quality of CL property tax records varied substantially by county and township in the U.S. On average, about 69% of 2010 ACS single-family, owner-occupied households could be linked to CL property tax data, but this rate was much higher in urban and affluent areas and much lower in rural ones. The likelihood of a record having CL information increased with household income, property value, and education level, and was sharply lower for households in poverty or with less education.

Dillon (2019) matched CL data to ACS household records in order to evaluate the possibility of using CL data values to substitute for ACS questions regarding acreage, number of rooms or bedrooms, tenure (rented or owned), property value, and property tax amount. Consistent with Seeskin (2016), Dillon (2019) found lower linkage rates among certain subpopulations. Dillon (2019) also found lower linkage rates for households with young children, minorities, residents of group quarters, low income residents, the unemployed, rural residents, and occupants with low education. Poorer match rates among these subpopulations could potentially introduce bias when using matched administrative data to impute missing values in the ACS. The overall linkage rate reported for CL to ACS household records was 64.2%. In conclusion Dillon (2019) found that using administrative records to replace or supplement selected ACS housing questions may be feasible for certain items (notably acreage), but issues of coverage, agreement, and conceptual differences mean most topics would be better suited for use in supplemental item imputation or validation rather than wholesale substitution.

Binder et al. (2022) matched AHS data to 2017 and 2019 CL data. Binder applied fuzzy matching to commercial records not having a MAF match via the production process. Binder found higher match rates among owner occupied units than renter occupied and among single

family structures than multi-unit buildings. As an added step, geospatial matching increased the match rate by 5% to 7%.

Both Binder et al. (2022) and Dillon (2019) found similar linkage rates between CL data and the MAF, ranging from 60% to 67%. Both researchers noted limitations in the potential use of commercial data for imputation.

More recently the Census Bureau evaluated BK property data in addition to CL data as a source for imputing or substituting for information reported in the ACS (Clark et al., 2023). BK is widely recognized for its comprehensive coverage of the U.S. real estate landscape. As with CL, the match quality between BK and ACS has also been rigorously evaluated. The study assembled a composite acreage file by linking ACS MAFIDs with BK property assessment and geospatial parcel boundary data from multiple years (2019 and 2021). Address-based linkage aligned parcel addresses to ACS housing units. Geospatial linkage joined housing unit coordinates to parcel polygons. Business rules screened out counties with poor administrative data quality (minimum 50% coverage and 80% agreement with ACS). The final BK-based acreage variable prioritized the most recent and reliable data, with geospatial data used selectively for single-family attached houses. In the terminology of Clark et al. (2023), a single-family attached house refers to a structure separated from adjoining units by a wall extending from the ground to the roof, for example, a townhouse or row house. Clark et al. (2023) found that about 87% of the full ACS sample of housing units could be matched to a unique property record in the BK dataset. Coverage was higher for certain structure types, with owner-occupied and single-family detached structures being easier to match than rentals or multi-unit buildings. Clark et al. (2023) concluded that given these match rates, it was feasible to use the BK acreage variable as a substitute for the ACS acreage variable.

4 Our Work

In this section we present our own preliminary work on developing and testing deterministic matching alternatives applied to the ACS and commercially available property data (BK) (Champney et al., 2024).

4.1 Problem Statement

We sought to improve the Census Bureau's production process to increase linkage rates of the MAFID to commercial property data. Increased matches would permit use of matched commercial property records to impute missing information or substitute for information in the ACS regarding property characteristics, such as assessed value, property tax, numbers of rooms and bedrooms, tenure, and the presence or absence of various characteristics such as a kitchen or bathroom. The BK data consisted of multiple files (basic address records, tax assessment records, and parcel boundary files), so one additional problem was how to best combine information from these multiple files when determining the best match. Since the BK data also included shape or property boundary files, we assessed how to best use these shape files in conjunction with the MAF geocodes to further increase match precision.

4.2 Methods

The data sources included 2019 ACS unswapped, unweighted housing unit records with MAFIDs attached via the Census Bureau's production process. We later applied the ACS weighting

scheme to attach sampling weights, but weighting does not affect the linkage algorithms themselves. The commercial side consisted of 2019 BK property data, including a base address file, tax assessment records, and parcel boundary (shape) files. All BK files had first been processed through the Census Bureau's standard MAFID assignment routine. This routine produced a baseline match rate of approximately 60% when compared to ACS housing units.

We summarize performance by housing type. For each housing type h, let N_h denote the total number of ACS housing units and M_{0h} the number of units matched in the production process. The baseline match rate for housing type h is defined as

$$B_h = 100 \times \frac{M_{0h}}{N_h},$$

the percentage of ACS units successfully linked in the baseline process. Each enhancement step produces additional matches beyond this baseline. For clarity, we report the relative improvement rates for each step as percentages of the baseline number of matches, rather than as percentage-point gains in the overall match rate. Specifically, if M_{1h}, M_{2h}, and M_{3h} denote the numbers of new matches obtained by Steps 1–3, the corresponding improvement rates are:

$$R_{1h} = 100 \times \frac{M_{1h}}{M_{0h}}, \quad R_{2h} = 100 \times \frac{M_{2h}}{M_{0h}}, \quad R_{3h} = 100 \times \frac{M_{3h}}{M_{0h}}.$$

These ratios express how many additional matches each method contributes, relative to the base production matches for that housing type. Because each R_{kh} is defined relative to the same baseline denominator M_{0h}, the three improvement rates are not additive, and the final match rate cannot be obtained by summing the baseline rate and the three improvement rates. Instead, the final match rate for housing type h is computed directly as

$$F_h = 100 \times \frac{M_{0h} + M_{1h} + M_{2h} + M_{3h}}{N_h},$$

the percentage of ACS units matched after all steps have been applied.

4.2.1 Step 1: Multi-File Deterministic Reconciliation

The BK data contained multiple files that covered the same underlying property universe (e.g., address, tax assessment, and parcel files), and the production process sometimes assigned different MAFIDs to the same BK property across these files. To assign a unique MAFID to each BK property, we implemented a deterministic voting and ranking procedure. For each BK property, we collected all candidate MAFIDs from the various BK files and assigned the MAFID that occurred most frequently ("voting"). When a tie remained, we broke ties using a pre-specified ranking of the BK files based on their expected address quality (for example, prioritizing the base address file over auxiliary files). This procedure reconciled inconsistent MAFID assignments and recovered additional matches beyond the original production process.

For each housing type h, let M_{1h} be the number of additional matches obtained through this multi-file reconciliation step. The relative improvement rate from Step 1 is

$$R_{1h} = 100 \times \frac{M_{1h}}{M_{0h}},$$

which appears in the "Improvement from multiple files" column of Table 1. Intuitively, R_{1h} indicates by what percentage the baseline number of matches increases when multi-file reconciliation is added.

4.2.2 Step 2: Geospatial Matching Using Parcel Boundaries

The second step used geospatial information to identify additional matches among records that remained unmatched after Step 1. Both the ACS and the MAF contain latitude and longitude coordinates for each housing unit, while the BK data include parcel boundary (polygon) shape files that delineate individual properties. Using the `sf` package in R (Pebesma and Bivand, 2023), we performed point-in-polygon operations to link ACS housing unit coordinates to BK parcel polygons at the county level. Working at the county level reduced memory demands and limited the search space for candidate matches.

Geospatial linkage can produce false positives, especially for multi-unit structures in which many housing units share the same geographic coordinates, or when parcel boundaries are coarse relative to the underlying units. We therefore applied a set of deterministic cleaning rules to the initial spatial matches. Specifically, we required agreement in house number between the ACS and BK records and, when either side indicated a multi-unit property, agreement in unit number. Spatial matches that failed these checks were discarded. We then added the remaining geospatial matches to the accepted matches from Step 1.

For housing type h, let M_{2h} be the number of additional matches contributed by Step 2 (over and above the baseline and Step 1 matches). The corresponding relative improvement rate is

$$R_{2h} = 100 \times \frac{M_{2h}}{M_{0h}},$$

reported in the "Improvement from geospatial linkage" column of Table 1. Again, R_{2h} reflects the percentage increase in matches relative to the baseline number of matches, not a percentage-point gain in the overall match rate.

4.2.3 Step 3: Fuzzy String Matching and Additional Cleaning

The third step applied fuzzy string matching to the remaining unmatched records to recover further links missed by both the production process and the geospatial step. We blocked the data by county to improve computational efficiency and reduce the risk of implausible cross-county matches. Within each county, we first constructed candidate pairs using simple blocking variables such as ZIP code and house number.

For each candidate pair of ACS and BK address strings, we computed three SAS edit-distance metrics: COMPGED, COMPLEV, and SPEDIS (Sloan and Hoicowitz, 2016). Before computing these distances, we removed punctuation and common stop words to focus the comparison on the core address content (street name, house number, and unit information). We examined the empirical distributions of each distance metric and their correlations and defined matches using a percentile-based rule rather than a single hard cutoff. Specifically, we required that all three distances fall within the most similar 25% of candidate pairs (i.e., below the 75th percentile of each distance). In addition, we imposed strict agreement on house number and, when present, unit number: we discarded candidate pairs with conflicting house or unit numbers regardless of their edit-distance scores.

We classified fuzzy-matching candidates that satisfied these criteria as accepted matches and added them to the set of links obtained from Steps 1 and 2. For housing type h, let M_{3h} be the number of additional matches contributed by Step 3. The corresponding relative improvement rate is

$$R_{3h} = 100 \times \frac{M_{3h}}{M_{0h}},$$

reported in the "Improvement from fuzzy matching and cleaning" column of Table 1. As with the other steps, we defined this rate relative to the baseline number of matches. It is not intended to be additive across steps.

We computed the final match rate for housing type h,

$$F_h = 100 \times \frac{M_{0h} + M_{1h} + M_{2h} + M_{3h}}{N_h},$$

directly from the total number of matched ACS units after all steps. It appears in the "Final match rate" column of Table 1.

4.3 Computation Costs

Costs associated with the matching methods consisted of computational time only. We conducted most of the computation on a server in the Census Bureau Integrated Research Environment (IRE) using SAS, which has the benefit of utilizing disk space instead of memory to process large datasets. SAS has an advantage over R or Python, which run in memory in the IRE, limiting the number of records that can be processed at a time to the available memory. For the geospatial matching we used the R sf package, in which case we ran the code at the county level sequentially for each county in the United States in order to accommodate the memory limitations, requiring more than about seven days of computing time to perform the matching on all of the counties.

4.4 Results

Table 1 summarizes linkage performance by housing type. The first column reports the baseline match rate, B_h (the percentage of ACS housing units linked in the Census production process). The next three columns report the relative improvement rates from each enhancement step: the number of additional matches generated by that step expressed as a percentage of the baseline number of matches for that housing type. Because each improvement rate uses the same baseline denominator, these three columns are not additive and do not sum to the final match rate. The final column reports the final match rate, F_h, defined as the percentage of ACS housing units matched after all three steps have been applied.

The baseline match rate B_h varies substantially across housing types, from 17.1% for multi-unit structures to 78.1% for single-family units. The relative improvement rates show how much

Table 1: Baseline match rates, relative improvements, and final match rates by housing type.

House type	Baseline B_h (% of ACS units)	Step 1 R_{1h} (% of baseline)	Step 2 R_{2h} (% of baseline)	Step 3 R_{3h} (% of baseline)	Final F_h (% of ACS units)
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

each enhancement step increases the baseline number of matches. For example, among trailer units the geospatial step alone adds 18.9% as many matches as the baseline production process, and the fuzzy step adds an additional 30.3% relative to baseline. In absolute terms, single-family units exhibit the largest percentage-point gain in final match rate, increasing from 78.1% to 88.1%, while "Other" housing types show a smaller absolute gain (from 19.3% to 25.5%) but a substantial proportional improvement over a low baseline. Multi-unit structures such as apartments and condominiums remain the most challenging, with a final match rate of 19.6% despite the additional matches contributed by all three steps.

5 Discussion

We found that each step we applied in sequence to matching provided incremental improvements in the match rate. The degree of improvement tended to vary by the type of housing structure, e.g., single family versus multi-unit dwellings. The poor match rate for multi-unit structures was consistent with other work by the Census Bureau (Brummet, 2014; Clark et al., 2023). Missing values and variation in the coding of housing unit numbers in multi-unit structures may be an important factor that attenuated match rates in these types of structures. Additional data sources with possibly better coding of unit numbers should be considered in future matching work.

Our multi-step, deterministic enhancement approach built upon the Census Bureau's standard production process by integrating cross-file reconciliation, geospatial joins using parcel boundaries, and fuzzy string matching. This provided a substantial relative gain in overall match rates—especially in regions and housing types where base rates were low. While the Census Bureau's matching routines rely heavily on production-standardized MAFID assignment and probabilistic matching (Brummet, 2015), our work demonstrates that targeted deterministic and geospatial refinement can recover true links missed in the initial pass, especially when leveraging available property boundary shapefiles (e.g., (Clark et al., 2023; Dillon, 2019). Consistent with existing Bureau analyses (Brummet, 2014; Binder et al., 2022)), our findings show highest matching rates for single-family homes and more limited success for multi-unit dwellings and nontraditional structures.

Both our work and prior Census Bureau research underscore persistent obstacles in accurately matching records for multi-unit buildings. Issues arise from missing or inconsistently coded unit identifiers and the use of parcel-level, rather than unit-level, reference data. Addressing these gaps will likely require new sources or improved field standardization. Our geospatial linkage using the R `sf` package (Pebesma and Bivand, 2023) and fuzzy string comparison (SAS COMPGED) (Sloan and Hoicowitz, 2016) aligns with a broader national trend—documented in Census studies (Clark and Sawyer, 2018; Clark et al., 2023)—of adopting spatial and approximate matching algorithms to supplement deterministic and probabilistic methods. Clark et al. (2023) and other researchers increasingly recognize that these hybrid approaches constitute a best practice for maximizing match rates while minimizing false positives. Improved linkage rates permit wider and more accurate use of administrative property data (e.g., Black Knight) for imputation or substitution of ACS variables such as acreage, assessed value, and property tax—directly supporting ongoing innovations in survey integration, as advocated in recent Census Bureau studies (Clark et al., 2023; Dillon, 2019; Binder et al., 2022). Achieving further gains will likely require investment in better unit-level data (especially for multi-unit dwellings), the adoption of robust hybrid or machine learning–based approaches (see e.g., Comber and Arribas-

Bel (2019); Zhang et al. (2024)), and continued monitoring of linkage quality by geography and housing type.

After applying sequential deterministic, geospatial, and fuzzy matching, a residual set of records is typically left unmatched—often due to complex errors, missing or highly variable data, or edge-case formatting (especially in multi-unit dwellings or areas with poor address standardization). Hagger-Johnson et al. (2017) applied probabilistic matching as a final pass to these records, enabling the capture of likely true matches that rigid or thresholded algorithms would miss.

Probabilistic methods (e.g., Fellegi–Sunter model (Binette and Steorts, 2022) as used in the Census Bureau's Person Identification Validation System) (Wagner and Layne, 2014) compute an overall match probability by combining the agreement or similarity of multiple address fields (e.g., street, city, ZIP, unit) with assigned weights. This allows the method to:

- Give more influence to reliable fields (e.g., house number, street name).
- Down-weight or tolerate minor disagreements or missingness in other fields.
- Identify matches with partial agreement, rather than requiring strict concordance.

By generating match scores (probabilities or likelihood ratios), probabilistic approaches provide a graded confidence level for each potential match. Follow-on work can:

- Set adaptive thresholds tailored to the dataset or subgroup (e.g., be stricter for multi-unit dwellings).
- Flag ambiguous or medium-probability pairs of records for clerical or expert review.
- Calibrate thresholds to balance precision and recall, depending on the operational needs.

Probabilistic matching could be layered upon the previous steps by using deterministically matched pairs of address records as training data to create a prediction model giving a match probability for each pair of records to be matched, spatial proximity could be incorporated into the modeling so as to give higher match probabilities to pairs of records more closely aligned geographically. Incorporating match probabilities is expected to help reduce false matches and identify true matched pairs of records missed by other methods. Probabilistic matching could be most beneficial as an added step for multi-unit and non-traditional structures. Probabilistic matching could also leverage additional housing unit attributes such as occupant and owner names and various housing characteristics. Recent Census Bureau methods articles, and working studies have found that probabilistic matching increased total link rates by 10–20% beyond deterministic and fuzzy routines alone, with manageable false match rates if thresholds were carefully tuned. These methods worked best when supplemented by quality truth data for calibration and after exhaustive rule-based and spatial matching stages (Clark et al., 2023).

In summary, our work underscores the value of iterative, hybrid matching strategies—carefully combining deterministic, geospatial, and fuzzy matching techniques—to maximize address linkage quality. These findings reinforce, extend, and operationalize key insights from the Census Bureau's own research agenda, pointing the way toward more comprehensive, equitable, and low-burden administrative record integration in federal statistics. Sequentially combining deterministic, geospatial, and fuzzy matching steps significantly enhances address linkage success rates—particularly for difficult-to-match properties and in counties with traditionally low administrative data coverage. While substantial gains are possible—as demonstrated by our 13.7% relative improvement—persistent weaknesses for multi-unit, rural, and non-traditional housing types remain a challenge for both our methods and existing Census Bureau routines.

Applying probabilistic matching as a follow-on step allows you to maximize match rates and data quality, especially for challenging records that resist deterministic or fuzzy approaches alone. By adding weighted, statistical logic to the matching sequence, one can recover more true

links, minimize false matches, and adapt to complex, real-world address data—enhancing the utility and reach of integrated datasets, in line with the Census Bureau's own evolving best practices (Clark et al., 2023).

6 Future Research

A key limitation of our current work is that we report improvements in overall match rates but do not directly quantify the accuracy of the resulting linkages, for example in terms of sensitivity, specificity, and related classification metrics. Future research should therefore implement a formal gold-standard validation study using the Census Bureau's ACS–MAFID production linkages as ground truth and evaluating our hybrid matching pipeline under controlled conditions.

One promising design would start with a set of ACS housing units to which Census Bureau MAFIDs were appended. A random subsample of these ACS records with attached MAFIDs (i.e., housing unit records) would be selected, stratified by key housing characteristics such as structure type (single-family, multi-unit, trailer, other) and geographic region. For this subsample, the MAFIDs would then be randomly masked, thereby simulating the situation in which the linkage identifiers are unknown while preserving the underlying address and geospatial information. The masked ACS records would then be re-linked to the MAF (and, where applicable, to Black Knight records) using the same deterministic, geospatial, and fuzzy matching sequence described in this article.

Because the original production MAFIDs are known but hidden during the re-linkage, they can be used as a gold standard for performance evaluation. For each ACS record in the validation sample, the re-assigned MAFID from the hybrid pipeline can be compared to the original production MAFID. This permits construction of a confusion matrix with the following components: true positives (records for which the re-assigned MAFID equals the original MAFID), false positives (records incorrectly linked to a different MAFID), false negatives (records that should have been linked but remain unmatched), and true negatives (records correctly left unmatched when no valid MAFID exists in the frame). From this confusion matrix, standard summary statistics such as sensitivity (recall), specificity, positive predictive value (precision), negative predictive value, and F1 scores can be computed for the overall pipeline or separately by housing type and region.

The design also naturally supports stepwise evaluation of each component of the hybrid matching algorithm. By recording, for each record, which step produced the first accepted match (multi-file deterministic consolidation, geospatial point-in-polygon linkage, or fuzzy string matching), one can estimate step-specific precision and recall relative to the gold-standard production MAFIDs. This would highlight, for example, whether geospatial matching introduces a non-trivial number of false positives in certain housing types, or whether fuzzy matching improves sensitivity for multi-unit dwellings at an acceptable cost in precision. In addition, varying the COMPGED distance threshold and spatial parameters (e.g., buffer sizes, coordinate system choices) within the validation sample would allow construction of empirical precision–recall curves, helping to select operating points that balance match quantity and quality.

Finally, the proposed gold-standard study could be extended to incorporate probabilistic matching as a final pass applied only to the subset of records remaining unmatched after deterministic, geospatial, and fuzzy steps. Using the masked ACS records with attached MAFIDs as truth, researchers could compare the incremental gains in sensitivity and F1 from adding probabilistic linkage, while closely tracking any degradation in specificity. Together, these eval-

uations would provide a rigorous quantitative assessment of hybrid matching strategies in the ACS–MAF context, directly addressing the need for accuracy and error-rate measures identified by the reviewers and guiding operational choices about threshold settings and workflow design in future Census Bureau linkage applications.

Supplementary Material

Zip file contains the original Powerpoint presentation, FCSM 2024 Matching Presentation Draft 10102024 Clean_CB.pptx and Word document describing files and matching process, JDS_matching_data_and_steps.docx.

Acknowledgement

The authors thank the editor, associate editor, and referees for their constructive comments which has led to significant improvement of this article.

Funding

This work was partially supported by the U.S. Census Bureau.

This work was supported by the U.S. Census Bureau, Department of Commerce (DOC) Contract 1331L523D13OS0003.

References

Aldridge RW, Shaji K, Hayward AC, Abubakar I (2015). Accuracy of probabilistic linkage using the enhanced matching system for public health and epidemiological studies. *PLoS ONE*, 10(8): e0136179. https://doi.org/10.1371/journal.pone.0136179

Binder A, Molfino E, Voorheis J (2022). Comparing the 2019 american housing survey to contemporary sources of property tax records: Implications for survey efficiency and quality, *Technical report*, U.S. Census Bureau, Center for Economic Studies.

Binette O, Steorts RC (2022). (almost) all of entity resolution. *Science Advances*, 8(12): eabi8021. https://doi.org/10.1126/sciadv.abi8021

Brummet Q (2014). Comparison of survey, federal, and commercial address data quality, *Technical Report CARRA Working Paper #2014-06*, U.S. Census Bureau, Center for Administrative Records Research and Applications.

Brummet Q (2015). Matching addresses between household surveys and commercial data, *Technical Report CARRA Working Paper #2015-04*, U.S. Census Bureau, Center for Administrative Records Research and Applications.

Champney T, Qin H, Coffey S (2024). Maximizing linkage in address data: Spatial, exact, and fuzzy matching. In: *Proceedings of the Federal Committee on Statistical Methodology Conference*. Federal Committee on Statistical Methodology. Accessed: 2025-08-07.

Clark S, Barth D, Binder A, Diagne AF, Pharris-Ciurej N, Voorheis J (2023). *Acreage administrative records data research for the american community survey Technical report*. U.S. Census Bureau.

Clark S, Sawyer R (2018). Housing administrative records simulation, *Technical report*, U.S. Census Bureau.

Comber S, Arribas-Bel D (2019). Demonstrating the utility of machine learning innovations in address matching to spatial socio-economic applications. REGION. *Journal of ERSA*, 6(2): 1–15.

Cortes TR, Silveira IHd, Junger WL (2021). Improving geocoding matching rates of structured addresses in Rio de Janeiro, Brazil. *Cadernos de Saúde Pública*, 37: e00039321. https://doi.org/10.1590/0102-311x00039321

Dillon M (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, *Technical report*, U.S. Census Bureau, Center for Economic Studies.

Dusetzina SB, Tyree S, Meyer AM, Meyer A, Green L, Carpenter WR (2014). An overview of record linkage methods. In: *Linking Data for Health Services Research: A Framework and Instructional Guide*. Agency for Healthcare Research and Quality (U.S.).

Goldberg DW, Wilson JP, Knoblock CA (2007). *From Text to Geographic Coordinates: The Current State of Geocoding. URISA Journal* **19**(1): 33–46.

Hagger-Johnson G, Harron K, Aldridge R, Fu B, Setakis E, ..., Gilbert R (2017). *Combining deterministic and probabilistic matching to reduce data linkage errors in hospital administrative data. International Journal of Population Data Science* **1**(1): 296. https://doi.org/10.23889/ijpds.v1i1.316

Lee S (2025). Mastering Shapefiles in Geospatial Analysis.

Moreno D (2025). The luca process: A local government's guide to the census address review.

Pebesma E, Bivand RS (2023). Spatial data science: With applications in R. In: *Handbook of Spatial Data Science*. Chapman and Hall/CRC.

Placekey (2024). Mastering address matching: Techniques and best practices. Industry blog/white paper.

Prindle J, Suthar H, Putnam-Hornstein E (2023). An open-source probabilistic record linkage process for records with family-level information: Simulation study and applied analysis. *PLoS ONE*, 18(10): e0291581. https://doi.org/10.1371/journal.pone.0291581

Raim A (2016). Informing maintenance to the U.S. Census Bureaus Master Address File using statistical decision theory.

Seeskin ZH (2016). Evaluating the use of commercial data to improve survey estimates of property taxes, *Technical report*, U.S. Census Bureau.

Sloan S, Hoicowitz D (2016). Fuzzy matching: Where is it appropriate and how is it done? SAS can help.

Tyris J, Dwyer G, Parikh K, Gourishankar A, Patel S (2024). Geocoding and geospatial analysis: Transforming addresses to understand communities and health. *Hospital Pediatrics*, 14(6): e292–e297. https://doi.org/10.1542/hpeds.2023-007383

US Census Bureau (2022). Chapter 3: Frame development.

US Census Bureau (2025). ACS information guide.

Wagner D, Layne M (2014). The person identification validation system (pvs): Applying the center for administrative records research and applications' (carra) record linkage software. *Technical report*. U.S. Census Bureau.

Zhang Z, Balsebre P, Luo S, Hai Z, Huang J (2024). *StructAM: Enhancing Address Matching through Semantic Understanding of Structure-aware Information*. In *Proceedings of LREC-COLING 2024*, 20–25 May 2024, pp. 15350–15361.

Zhu Y, Matsuyama Y, Ohashi Y, Setoguchi S (2015). When to conduct probabilistic linkage

vs. deterministic linkage? A simulation study. *Journal of Biomedical Informatics*, 56: 80–86. https://doi.org/10.1016/j.jbi.2015.05.012