

©Copyright 2026

Suyun Catherine Kim

Geocoding Pipeline Development and Data Quality Assessment
For Enhancing Precision Medicine Research

Suyun Catherine Kim

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2026

Committee:

Dr. Jimmy Phuong, MSPH, PhD

Dr. Bryant T. Karras, MD, FACMI

Dr. Stephen J. Mooney, MS, PhD

Program Authorized to Offer Degree:

Biomedical Informatics and Medical Education

University of Washington

Abstract

Geocoding Pipeline Development and Data Quality Assessment

For Enhancing Precision Medicine Research

Suyun Catherine Kim

Chair of the Supervisory Committee:

Jim Phuong

Biomedical Informatics and Medical Education

Geographic location shapes health outcomes through environmental exposures, socioeconomic conditions, and access to care. However, the quality of geocoded address data to facilitate these inferences remains poorly characterized to understand the positional error from geocoding nationwide. This thesis presents a geocoding benchmarking study evaluating the DeGAUSS geocoding framework across a curated reference dataset of over 270,000 public locations spanning U.S. states and territories. Address preprocessing applied conformance and formatting checks aligned with USPS Publication 28 standards, including flags for embedded city or state information, malformed unit identifiers, missing spaces, and state-specific address patterns. Using verified GIS coordinates as reference coordinates, positional error between the reference coordinate and the geocode output was calculated via the Haversine formula and evaluated across five DeGAUSS match score thresholds (60%–100%), with territories analyzed separately due to substantially higher error and poor reference coordinate quality. Principled outlier removal

filters out geocodes that fall outside the expected state boundary, distinguishing misclassification (geocoder error) from misannotation (reference coordinate error). Relaxing the match score threshold from 100% to 60% increased successfully geocoded records from 58.3% to 86.9% of the dataset, while producing only modest increases in median positional error nationally from 65.3m to 68.0m. In states with higher error, lower-scoring records showed meaningfully worse positional error. Geographic stratification revealed substantially lower performance in U.S. territories, where median positional error exceeded 450m even in Puerto Rico and surpassed 4,000m in Guam. Together, these findings suggest that geocoding quality standards may need to incorporate geographic context to support accurate residential linkage in integrative precision medicine research.

Table of Contents

ACKNOWLEDGMENTS.....	8
INTRODUCTION.....	9
The Impact of Geographic Location on Health.....	9
Geocoding and Public Health.....	10
Geocoding Process.....	11
Geocoding Landscape.....	12
All of Us Research Program - Enrollment Address History.....	14
METHODS.....	15
CLAD Geocoding Pipeline.....	15
Preprocessing For Address Field Conformance and Formatting Checks.....	16
Preprocessing For DeGAUSS Inputs.....	19
DeGAUSS Containerized Geocoding.....	19
Post Hoc Analysis Geocoding Iteration 1.....	22
Address Correction for Detected Address Anomalies.....	24
Post Hoc Analysis Geocoding Iteration 2.....	26
Reference Data Description.....	27
Outlier Exclusion.....	28
Rural, Native Area, and Territory Classification.....	29
RESULTS.....	30
DISCUSSION.....	44
Geocoding Quality and the Choice of Success Threshold.....	44
Defining Success: How Match Criteria Compare Across Studies.....	46
Precision Level and Geomarker Linkage.....	46
Geographic Variation in Geocoding Performance.....	48
Implications for Environmental and Precision Medicine Research.....	49
Infectious Disease Surveillance and Public Health Applications.....	50
Distinguishing Geocoder Error from Reference Data Error.....	50
CONCLUSION.....	51
REFERENCES.....	54

Table of Figures

Figure 1. Overview of CLAD geocoding process.....	15
Figure 2. Overview of Preprocessing For Address Field Conformance and Formatting Checks.....	16
Table 2. Summary table of flags in the preprocessing output.....	17
Figure 3. Overview of Post Hoc Analysis Geocoding Iteration 1 and Geocoding Iteration 2.....	22
Table 3. Summary of principled outlier removal method.....	28

Table 4. Geographic stratification summary comparing urban vs rural vs tribal vs territory DeGAUSS geocoder performance following principled outlier removal.....	30
Table 5. Classification of geocoding outcomes by reference coordinate accuracy and geocoded coordinate accuracy, based on principled spatial validation against 2025 Census TIGER/Line state boundaries.....	31
Table 6. Positional errors of low-error state (California), a mid-range state (Washington), and a high-error state (Alabama) based on precision of range geocodes after principled outlier removal.....	32
Figure 4. Choropleth map of median positional error by state, showing geographic variation across the U.S. (Puerto Rico is excluded from the color scale) based on precision of range/street.....	33
Table 7. Geocoding positional error by state among states with $\geq 1,000$ successful geocoding records (precision of range or street, score ≥ 0.8 , principled outlier removal). Error was measured as straight-line distance in meters between geocoded and reference coordinates, and states were sorted by median error, descending.....	33
Figure 5a. Boxplots illustrating the within-state positional error variation and the contrast between a low-error state (California), a mid-range state (Washington), and a high-error state (Alabama). Whiskers extend to $1.5 \times \text{IQR}$ beyond the 25th and 75th percentiles.....	36
Figure 5b. Boxplots illustrate positional error by match score threshold for California, Washington, and Alabama.....	37
Figure 6. Positional error distributions for three representative states at match score thresholds of 100%, 90%, and 80% (precision of range only, principled outlier removal). All panels share a common log-scale y-axis. Whiskers extend to $1.5 \times \text{IQR}$ beyond the 25th and 75th percentiles.....	39
Table 8a. Threshold sensitivity summary for all states (precision of range only).....	40
Table 8b. Threshold sensitivity summary for California (precision of range only).....	41
Table 8c. Threshold sensitivity summary for Washington (precision of range only).....	41
Table 8d. Threshold sensitivity summary for Alabama (precision of range only).....	41
Figure 7. Percentage of range-precision geocodes by positional error bin and match score threshold, after principled outlier removal (U.S. states only).....	42
Table 9. DeGAUSS precision level distribution for valid geocodes after principled outlier removal (n=270,445).....	43
Table 10. Individual territory results with sample size, median error, and IQR.....	43

ACKNOWLEDGMENTS

I would like to thank my committee chair, Dr. Jimmy Phuong, and committee members Dr. Bryant T. Karras and Dr. Stephen J. Mooney for their guidance throughout this work. I am also grateful to the fellow lab group members in jimscoolnewgroup and the Center for Linkage and Acquisition of Data (CLAD) Geospatial team for their collaboration and support.

INTRODUCTION

The Impact of Geographic Location on Health

Beyond healthcare access alone, geographic location can shape health through a broad range of environmental and structural factors. The walkability of a neighborhood, the availability of green space, exposure to air pollution, proximity to fresh produce markets, and the overall cost and safety of one's environment all can contribute to health outcomes. For instance, Mitchell et al. (2008) found that populations with the highest exposure to green environments showed the lowest levels of health inequality related to income deprivation.¹ Moreover, biological sensitivity, socioeconomic conditions, and geographic location each shape vulnerability to climate-sensitive health outcomes, such as heat stress, air pollution effects, and water-, food-, and vector-borne illnesses. Those at greatest risk for these climate-sensitive outcomes and related infectious diseases include children, pregnant women, older adults, individuals living in poverty or with chronic conditions, outdoor workers, and communities in coastal or low-lying riverine areas.²

Altogether, these factors underscore why understanding geographic disparities in health is important. At the individual level, where a person lives can be as predictive of their life expectancy as their genetic makeup or lifestyle behaviors.³ At the population level, identifying geographic clusters of poor health outcomes allows public health officials and policymakers to design interventions where they are needed most and allocate healthcare resources more equitably. Thus, accounting for geographic location and variation is critical for public health research aiming to improve the health of vulnerable communities.

Geocoding and Public Health

Geocoding is a process that involves mapping a location's description, like an address, to its geographic coordinates on the earth. These true locations, in the form of a pair of coordinates or geographic features with attributes, can then be used for mapping or spatial analysis,⁴ which can be useful in public health applications like environmental exposure assessment,⁵ epidemiological research,⁶ and health accessibility research.⁷ More broadly, geocoding is a foundational tool for identifying vulnerable populations, evaluating the accessibility of care across geography, and detecting hotspots of poor health outcomes such as areas with elevated infectious disease risk or chronic air quality exposure events. By identifying where adverse outcomes cluster repeatedly over time, geocoding enables researchers to identify recurrent risk factors and inform broader public health policy and action.

The spatial and temporal units of analysis used in public health geocoding research vary considerably depending on the research question. Spatially, studies have employed units ranging from fine-grained individual address points and census blocks to census tracts, ZIP codes, counties, and broader regional or national boundaries.⁸ Finer spatial resolutions like individual addresses can be used to link a pharmacy with a medication delivery service to a patient living in an area with few pharmacies,⁹ while larger units can help identify neighborhoods with children with increased morbidity and inform targeted interventions.¹⁰

Accurate and standardized geocoding allows researchers to connect participants to neighborhood characteristics, pollution exposures, built environment features, climate risks, and area-level socioeconomic indicators. When health outcomes are sensitive to the magnitude of exposures in a geographic location, any loss in geocoding accuracy can make it more difficult to establish relationships between those exposures and health outcomes.⁵ When addresses are

incorrectly recorded, misspelled, or assigned the incorrect geocode,¹¹ geocoding methods can vary in accuracy due to differing resolutions of the reference data and methods for linking addresses.¹² Understanding how geocoding performance varies across methods, geographies, and quality thresholds is important for improving address data quality and ensuring the validity of environmental and public health research. The validity of spatial analyses depends on the accuracy and appropriateness of the geocoding method used.

Geocoding Process

To assess geocoding outcomes, we can evaluate their precision (how close the geocode is to the true location), certainty (how confident are we that the geocode represents the true location), and reliability (how much do we trust the geocoding method to create a geocode).¹³ Common metrics include match rates, match type, match score, spatial accuracy, and administrative unit concordance (Table 1).

Table 1. Geocoding system quality metrics as described in Goldberg et al. (2013),¹⁴ Mazumdar et al. (2008),⁵ Han et al. (2013).¹⁵

Source	Quality Metric	Description
Goldberg et al. (2013)	Match rate (%)	Percentage of all records capable of being geocoded
	Match type (% by geographic level)	Geographic levels of geocode match – building level, parcel level, street centroid level, postcode level, etc. and percentages of matchable geocodes at each level
	Match score (% at score levels)	Continuous similarity quantity of agreement between the input address and the matched reference address

	Spatial accuracy (% at distance levels)	Frequency distribution of distances between matchable geocodes and ground truth locations
	Administrative unit concordance (% variation from other systems)	Frequency distribution of administrative unit concordance between the same geocode produced by multiple geocoding systems.
Mazumdar et al. (2008)	Geocoding error	Difference in distance units between geocode and reference geocode for a given address
Han et al. (2013)	Spatial clustering analysis	Identifies geographical areas where the proportion of non-geocoded addresses was higher than expected

Geocoding Landscape

Many studies have examined geocoding accuracy using a range of geocoding methods and quality assessment techniques. Mazumdar et al. (2008) assessed geocoding accuracy to compare different geocoding methods through metrics like positional accuracy and comparison to the results of orthophoto map-based geocoding as a reference.⁵ They collected residential addresses, environment data of a disease-causing contaminant source called CAFOs (Concentrated Animal Feeding Operations), and simulated health data from Carroll County, Iowa.⁵ They then applied TIGER geocoding (automated matching based on street topology) and E-911 geocoding (automated matching to parcel or property location) methods to this data.⁵ To assess the accuracy of these methods, they looked at their positional accuracy (spatial displacement between the geocoded address locations and their reference location), match rate (proportion of addresses successfully geocoded with high confidence), and positional deviation (straight line distance between the reference and alternative geocodes).⁵ After simulating probabilities of disease, Mazumdar et al. found that geocoding accuracy significantly affects the

ability to reliably detect and estimate relationships between environmental exposures and health outcomes.⁵

Faure et al. (2017) also assessed positional accuracies of different geocoders to compare them.⁶ They collected address records representing the residential history of 1,685 women from Rhône-Alpes region in the French national E3N cohort analyzing the association between environmental dioxin exposure and breast cancer risk.⁶ Next, they utilized 3 different geocoding methods including a reference method, and each provided different categories of accuracy levels.⁶ They compared positional accuracies across these categories, used Chi square tests to assess statistical differences in accuracy levels, and performed Kappa statistics to assess agreement of positional accuracy of the geocoding methods.⁶ From these metrics, Faure et al. found that geocoding accuracy was similar in the two non-reference geocoding methods, but geocoding accuracy was significantly better in urban areas compared to rural areas.⁶

One geocoding framework and software tool designed specifically for epidemiologic and public health research is DeGAUSS (Decentralized Geomarker Assessment for Multi-Site Studies).¹⁶ DeGAUSS is open-source geocoding software that allows researchers to perform geocoding and geomarker linkage locally using containerized workflows.¹⁶ This decentralized approach keeps address data in secure environments while still allowing standardized spatial processing across research sites.¹⁶ In addition to geocoding addresses, DeGAUSS can link environmental, neighborhood, and contextual exposure variables to residential locations, making it useful for large multi-institutional studies that require reproducible spatial data processing.^{16,17}

All of Us Research Program - Enrollment Address History

The National Institute of Health All of Us Research Program is a nationwide precision medicine project across the United States and Territories, offering a level of geographic diversity.¹⁸ Address data within the program are collected through surveys administered at patient enrollment, providing a participant pool that is nation-wide. There have been national-scale studies that routinely perform geocoding and linkage for place-based exposures.¹⁹ Systematic reporting of geocoding quality metrics across the U.S. states and territories remains scarce. Estimation of positional error across US geographies are needed to establish confidence in environmental data linkage, like for the All of Us Research Program.

This thesis describes a geocoding benchmarking analysis conducted in collaboration with the National Institute of Health All of Us Research Program Center for Linkage and Acquisition of Data (CLAD). Here, we describe the conceptual flow of the CLAD Geocoding Pipeline and the post hoc analysis performance of the DeGAUSS geocoder using a curated dataset of over 270,000 public reference locations spanning U.S. states and territories from the Homeland Infrastructure Foundation-Level Data (HIFLD) repository.²⁰

A central research question driving this work is: how do definitions of geocoding success change the composition and positional accuracy of the resulting dataset? We address this through a sensitivity analysis examining (1) the quality thresholds at which geocodes are deemed high-quality and successful, and (2) the change in positional error among geocodes classified as successful when that threshold is relaxed. Answering these questions would strengthen guidance for applying DeGAUSS and similar interpolation-based geocoders to large, geographically diverse datasets, and would support more accurate linkage of individuals to environmental, social, and structural determinants of health.

METHODS

CLAD Geocoding Pipeline

The CLAD geocoding process occurs in 5 phases: (1) preprocessing and residential history cleaning, (2) geocoding iteration 1, (3) post hoc analysis, (4) geocoding iteration 2, and (5) reconciling the best geocode across runs for delivery to the Environmental Data Linkage team (Figure 1). Geocoding iteration 2 and subsequent runs include iterative post hoc analyses to appraise the geocoded results. The preprocessing steps (Figure 2) and post hoc analysis described in subsequent sections were developed as part of this pipeline and inform the quality assessment framework applied in this benchmarking study.

Figure 1. Overview of CLAD geocoding process.

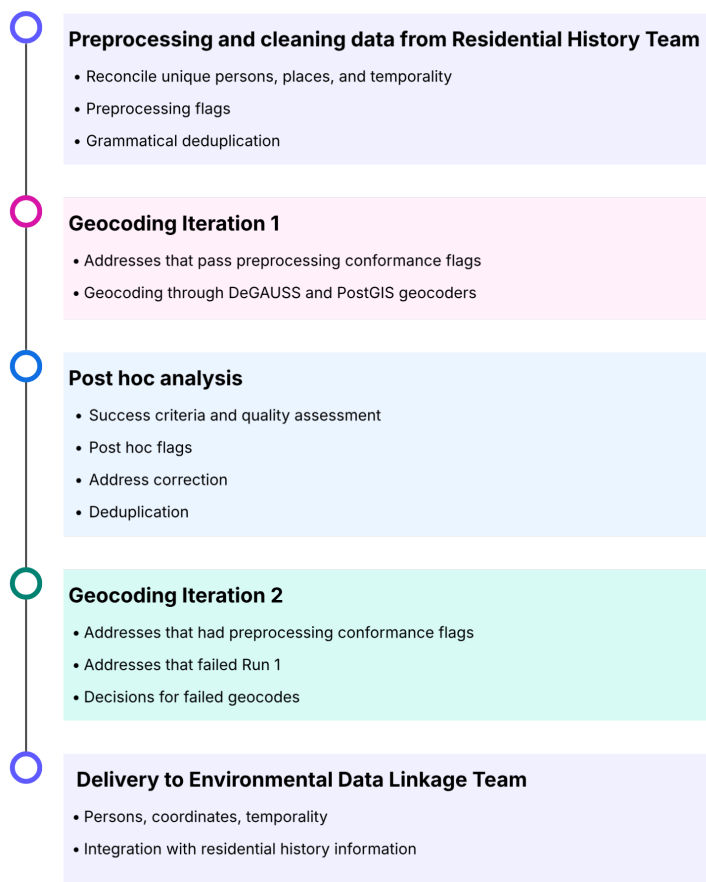
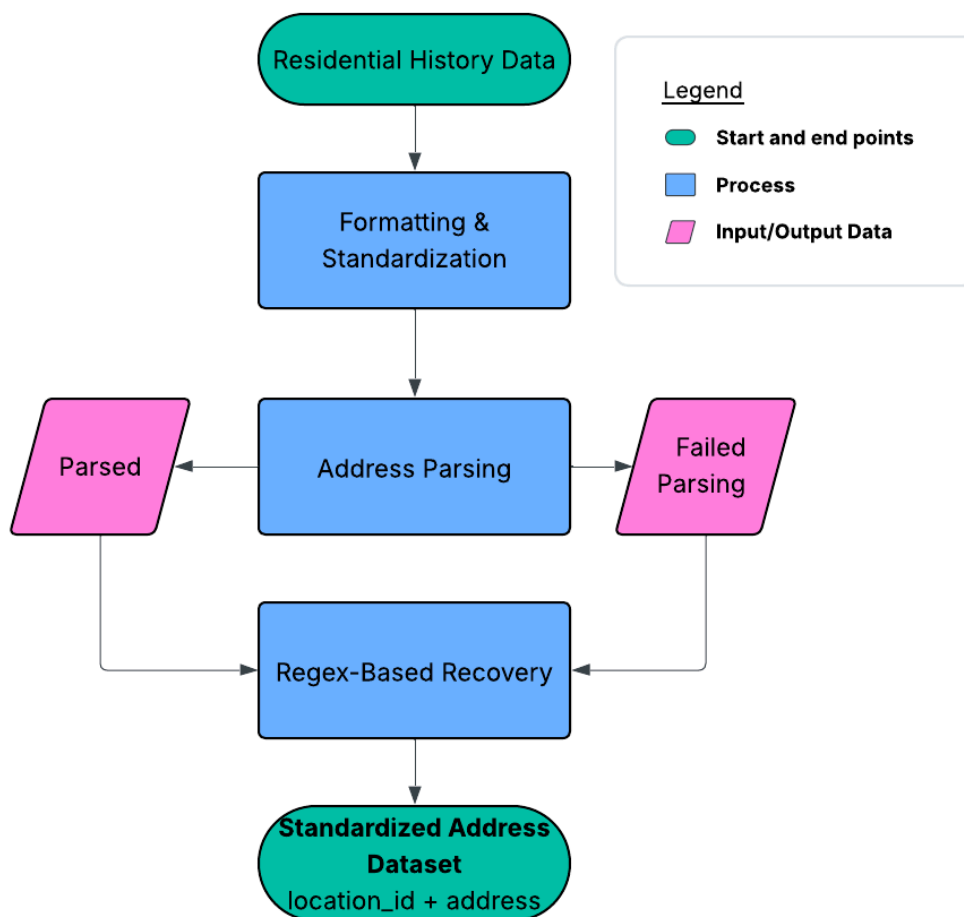


Figure 2. Overview of Preprocessing For Address Field Conformance and Formatting Checks.



Preprocessing For Address Field Conformance and Formatting Checks

The goals of preprocessing are 1) to standardize the address components to USPS Publication 28 Standard format and 2) join the address components into a single address string for geocoder input (<address_1>, <city>, <state> <zip>).²¹ In collaboration with the CLAD residential history team, the preprocessing phase of the Geocoding pipeline involved defining mapping requirements to preserve linkages across person, place, time, and external data sources. This included generating an internal person_id for each person of interest, a location_id for each place of interest unique to each person, and incorporating temporal information for each

residential record. Location identifiers were assigned for each individual to account for situations with residence subunit details like apartment and building numbers or shared household information that was incomplete or missing. The existing preprocessing pipeline starts by reading the residential history record input and applying formatting checks using custom flagging functions (Table 2). These functions use regular expressions to recognize common structural characteristics issues in the street address field observed from a sample of the CLAD enrollment address history data. Common errors detected in the address_1 field include embedded or repeated city or state information, missing spaces after street numbers, or improper apartment or unit formatting. Each record is evaluated with these flags, and rows with any detected formatting issue or duplicated location ids are removed. Addresses with flags are reviewed and corrected prior to geocoding. This creates a cleaned location history table containing standardized full address strings for each record for downstream DeGAUSS geocoding and data linkage.

Next, the location history dataset enters a parsing stage in which structured components from the full address string including address_1, address_2, city, state, and zip are put into a new OMOP-formatted location table while preserving the original location_id and format_check fields. Additionally, invalid placeholder string values are replaced with NaN, and state and territory names are standardized into postal abbreviations, and the structured address components are then written into a new Foundry dataset.

Lastly, the preprocessing script finds records that were not successfully or cleanly parsed due to formatting errors. These records are passed into a custom regex-based parsing function that uses regex patterns and additional logic to extract state abbreviations, full state names, and zip codes to recreate structured components for addresses.

Table 2. Summary table of flags in the preprocessing output.

Flag Column Name	Description/Purpose
Address1 anomaly feature flags	
flag_for_unit_apartment_number_removal	Flags if a unit or apartment number is found in the address_1 field but should be relocated to the address_2 field.
flag_for_PO_BOX_numbers	Flags if the address contains a PO BOX number (special handling may be required).
flag_for_missing_space_after_stno	Flags if there is a missing space after the street number in the address_1 field.
flag_for_predirection_space_fix	Flags if there is a missing or extra space in the predirectional part of the address_1 field.
flag_for_space_between_street_abbrev	Flags if there is a missing or extra space between the street name and its abbreviation.
flag_whole_embedded_address	Flags if the address_1 field has additional content embedded within the entry, like city, state, or zip code (potential parsing issue from extraneous information).
flag_grouped_addr1	Flags if the street address contains grouped or concatenated elements that should be separated with proper spacing and delimiters.
flag_possible_apartment	Flags if the address_1 ends with an orphaned value that could be an apartment/unit indicator.
flag_extra_spaces_content	Flags if there are extra spaces within the address content (potential formatting and parsing issue).
Regionally specific address format	
flag_for_IL_WI_alnum_grid_stno	Flags for a regional address format with alphanumeric grid street numbers (special format found in Illinois or Wisconsin).
flag_for_NY_CA_HI_hyphenated_addr	Flags for addresses with hyphenated street numbers (special format in New York, California, or Hawaii).

During this phase, modifications were made to the preprocessing script while revising the workflow. An error occurred when converting a Pandas DataFrame to a Spark DataFrame because newly created columns were initially populated with Python type objects (int, str). Since Spark infers column data types from actual values, it could not infer a valid schema from these objects. To resolve this issue, placeholder values were replaced with `numpy.nan`, which

represents missing data and allows Spark to infer appropriate column types. Additionally, columns not included in the intended output schema were removed prior to conversion to ensure compatibility with the Spark DataFrame structure.

Preprocessing For DeGAUSS Inputs

For geocoding preparation, the preprocessing output dataset was read and reduced to the fields required by the DeGAUSS geocoder. The dataset retained unique location identifiers (`location_id`) and standardized full address strings, while all other metadata fields were removed. The address field was renamed to match the input schema expected by the DeGAUSS container. The resulting dataset contained `location_id` to preserve record linkage and the renamed address column formatted as `<address_1>`, `<city>`, `<state>` `<zip>`. To prepare the data for containerized geocoding, this dataset was exported as a comma-separated values (CSV) file containing the `location_id` and address fields, consistent with the standard input format required by the DeGAUSS geocoder.¹⁶

DeGAUSS Containerized Geocoding

Standardized address strings were geocoded using the DeGAUSS geocoder v3.3.0, which was deployed as a Docker container.²² Docker containers package software together with its required dependencies in a portable runtime environment, allowing the geocoder to be executed consistently across computing systems.²³ In this workflow, the DeGAUSS container was executed from the command line with the input CSV file mounted into the container environment, where the geocoder read the address column and processed the address records in batches within the containerized environment.

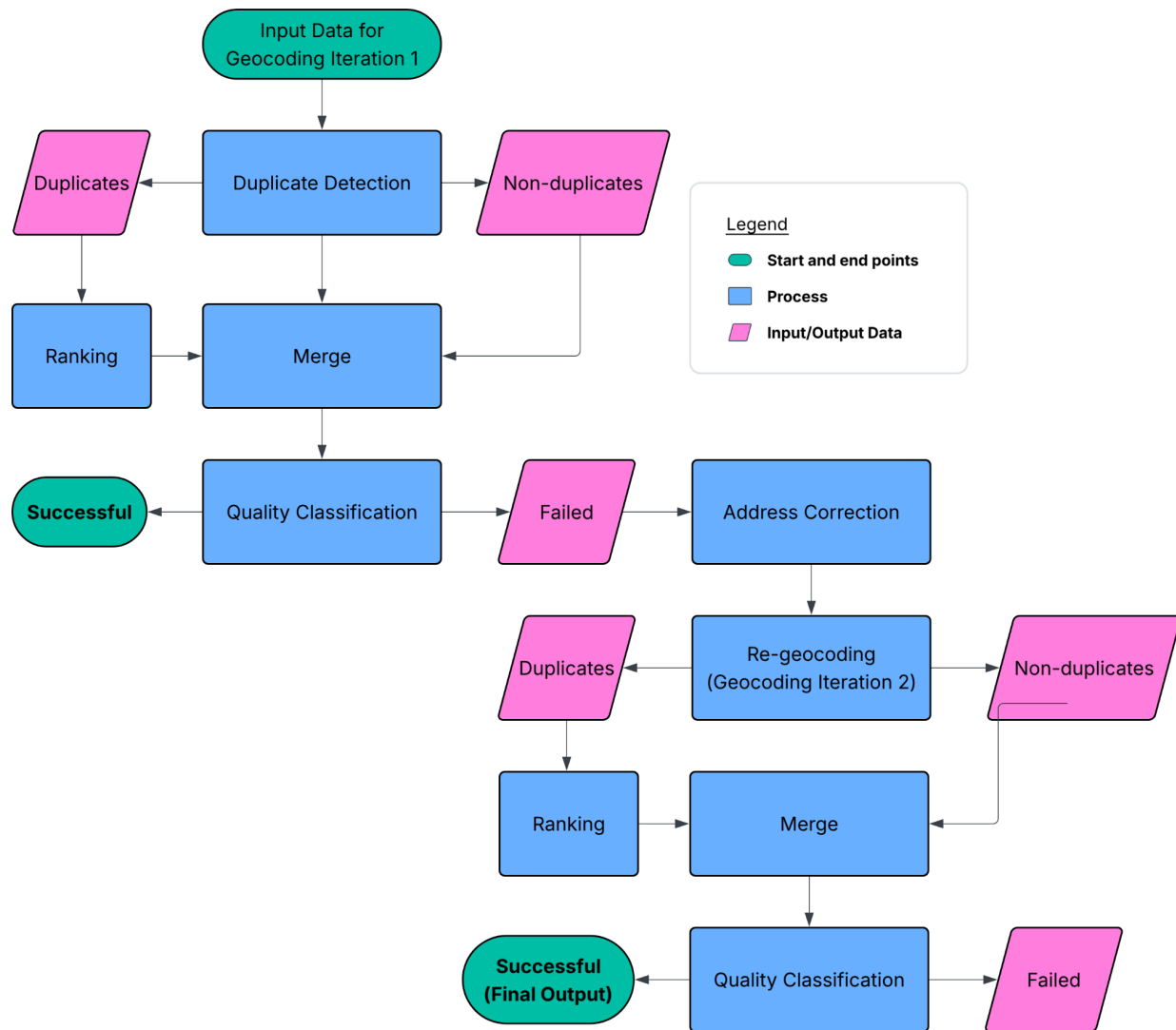
First, the geocoder performs address parsing by matching the input address string against a reference address database. Through this process, matched address components are extracted, including `matched_street`, `matched_city`, `matched_state`, and `matched_zip`, which correspond to the best reference match for each record.

Second, each record is assigned a precision level that reflects the spatial resolution of the geocoded result. Precision categories include `range` (address-range interpolation along a street segment and representing the highest level of precision), `street` (street centerline match), `intersection` (intersection-level match), `zip` (ZIP code centroid), `city` (city centroid), and `nan` (failed geocode). In addition, the geocoder calculates a similarity score on a 0–1 scale that quantifies the textual agreement between the input address string and the matched reference address components with higher scores indicating stronger matches. Since scores are calculated relative to the assigned precision category and length of address string used, they are interpreted within precision levels rather than compared across different precision types. Based on the assigned precision level and score, latitude and longitude coordinates (`lat` and `lon`) are generated to represent the matched street address range and interpolated location. Interpolated coordinates (from a precision level of `'range'`) may correspond to census blocks using address range (`ADDR`) information along a street between two cross streets (e.g., `'150 main street'` is on `'main street'` and may be interpolated using the street segment within 100-200 address range). To reduce the use of low-confidence geocodes, the geocoder suppresses latitude and longitude coordinates for results with precision below street level. In accordance with the default configuration used in this workflow, coordinates are suppressed for geocodes with precision below street level or a match score below 0.5.

Finally, the geocoder identifies special address types that are not eligible for standard geocoding. These include post office box addresses (`po_box`), institutional or foster care addresses (`cincy_inst_foster_addr`), and invalid or non-address text strings (`non_address_text`). These classifications are summarized in the `geocode_result` field. The container runs produce a log file documenting execution status and geocoding performance, which supports monitoring and reproducibility of the geocoding workflow. For posthoc sensitivity analyses, we applied two thresholds to evaluate geocode quality. The first analysis included only geocodes with a minimum precision of range and a semantic score of at least 0.8, while the second analysis included only geocodes with a minimum precision of range and a semantic score of at least 0.7, or a precision of street with a semantic score of at least 0.8.

Post Hoc Analysis Geocoding Iteration 1

Figure 3. Overview of Post Hoc Analysis Geocoding Iteration 1 and Geocoding Iteration 2.



Following DeGAUSS geocoding iteration 1, a structured post hoc analysis pipeline was implemented to evaluate match quality, correct address inconsistencies, and prepare unresolved records for subsequent reprocessing (Figure 3). The Input dataset included successfully preprocessed addresses and flagged records from earlier quality checks, which were linked to person-level residential histories and merged with raw DeGAUSS outputs containing coordinates, match precision, and similarity scores. Initial quality assessment focused on

identifying duplicate location identifiers and quantifying overlap between successfully geocoded records and those requiring further review.

Geocoding quality was assessed using composite ordered score (`ord_score`), which combined categorical precision levels (e.g., range, street, city) with the continuous DeGAUSS match score to prioritize the most reliable results. Standardized pseudo-address strings were generated to enable direct comparisons between input and matched addresses. A series of targeted rule-based corrections was then applied to fix common formatting artifacts, including removal of extraneous text, normalization of misencoded street terms, and extraction of secondary address components like apartment or floor numbers.

Duplicate records were resolved by selecting the highest-quality geocode per location based on the composite score. Records were then classified into success and failure categories using predefined thresholds that incorporated both precision and match score, with additional distinctions for corrected matches and partial failures. At the same time, discrepancies between input and matched address components were flagged to identify potential geocoding inaccuracies or input data errors.

A comprehensive address cleaning and standardization process followed, involving both general and region-specific rules. These included correcting spacing issues, separating concatenated address elements, standardizing directional and street abbreviations, and handling special cases such as alphanumeric street numbers or hyphenated formats common in specific U.S. regions. Unit and PO Box information were systematically extracted into secondary fields to improve geocoding compatibility.

City–state–ZIP (CSZ) consistency was then validated against external reference datasets, enabling detection and correction of mismatched geographic identifiers. For instance, if there

was sufficient agreement between the city and ZIP, the state was corrected accordingly. All flags, corrections, and geocoding outcomes were consolidated into a unified dataset, which was subsequently partitioned into successful and failed subsets. Successfully geocoded records were retained for downstream spatial analyses, while failed records annotated with detailed error flags were routed into a second geocoding iteration. Finally, summary statistics were generated to quantify geocoding performance, including success rates, duplicate resolution outcomes, and geographic variation in match quality.

Address Correction for Detected Address Anomalies

A post-hoc address correction and validation pipeline was implemented to improve geocoding success following geocoding iteration 1. This process systematically identified, corrected, and validated problematic addresses excluded from geocoding iteration 1 or not successfully processed in geocoding iteration 1 and therefore needed additional cleaning.

The correction workflow combined targeted fixes with broader rule-based transformations. Known formatting issues such as misencoded street suffixes, extraneous parenthetical information, duplicated terms, and special characters were corrected using standardized string operations. Street type abbreviations were normalized, and common errors in numeric and ordinal representations were resolved. Flags identified structural address issues, including embedded city or state information within address lines, grouped or concatenated addresses, missing spaces, misplaced unit or apartment identifiers, PO Box formatting inconsistencies, and state-specific address patterns (e.g., grid-style or hyphenated formats).

Flagged records were subsequently processed using corresponding correction functions tailored to each issue type. These transformations included parsing and relocating unit

information to secondary address fields, separating multiple addresses, inserting missing delimiters, and removing invalid characters. Regionally specific logic was incorporated to account for known variations in address conventions across states.

City–State–ZIP combinations were further validated against external reference data. Addresses were evaluated as triads, and inconsistencies were identified by comparing observed combinations to known valid mappings. When two of the three components were consistent, the third was corrected accordingly. Additional normalization addressed common city abbreviations (e.g., “PHX” to “PHOENIX,” “ATL” to “ATLANTA”) and known edge cases through targeted rules. Records that could not be corrected due to missing a valid street number, representing non-residential, incomplete entries to name a few were flagged. The final output preserved both original and revised address components, along with all issue flags and indicators of applied corrections.

Additional modifications were made to the flagging functions to improve detection sensitivity and the ability to handle inconsistent or imperfect address data. Specifically, changes to the `flag_whole_embedded_address` logic addressed limitations in the original version, which only executed when the city field was non-null and required an exact string match between the city and address line. This approach failed to capture cases involving misspellings or minor variations. The updated version introduced fuzzy matching to allow approximate string comparisons, incorporated case normalization to ensure consistency regardless of capitalization, and added null checks to prevent runtime errors. Modifications were also made to the `flag_grouped_addr1` function to include null handling and replace strict regex-based matching with fuzzy matching logic. These improvements increased the flexibility and accuracy of pattern detection for concatenated or grouped address entries. Collectively, these modifications

strengthened the reliability of the flagging system, enabling more comprehensive identification of address anomalies and supporting more effective downstream corrections.

Post Hoc Analysis Geocoding Iteration 2

Geocoding iteration 2 extended the post hoc framework established in geocoding iteration 1 by incorporating both previously flagged records and newly processed addresses. Input datasets included flagged outputs from geocoding iteration 1, geocoding iteration 2 query results with applied corrections, raw DeGAUSS geocoding outputs, and person–location linkage data. Records from both runs were merged and standardized, with duplicate location identifiers identified and resolved using a composite scoring metric that combined geocoding precision and similarity scores to prioritize the highest-quality match.

Geocoding quality was reassessed using refined success criteria, with thresholds based on match precision and similarity scores. Records were categorized into successful, successful with corrections, and failed groups, with particular attention to discrepancies between input and matched address components. Concurrently, CSZ validation was performed against external reference datasets to identify invalid or inconsistent geographic combinations. Where agreement between two components was established, the remaining component was corrected using reference-based lookups.

Final outputs were partitioned into three datasets: successfully geocoded records suitable for downstream spatial analysis, failed records requiring further review or correction, and a comprehensive flagged dataset containing all intermediate flags, fixes, and geocoding outcomes. As an additional quality assurance step, spatial validation was conducted by intersecting

geocoded coordinates with ZIP Code Tabulation Area (ZCTA) boundaries to assess geographic consistency.

Reference Data Description

The benchmarking dataset used in this study was curated by the CLAD Geospatial team from the Homeland Infrastructure Foundation-Level Data (HIFLD) repository, a publicly available collection of geospatial datasets maintained by the Department of Homeland Security.²⁰ Multiple shapefiles representing public and commercial buildings across the U.S. were downloaded, including hospitals, FDIC-insured banks, public schools, and places of worship, to achieve adequate geographic coverage across states and territories. For each location, a full address string in the format of street address, city, state, and zip code was constructed, and a geographic centroid was calculated from each polygon feature. All records were incorporated into a single dataset.

Address components were parsed using the `usaddress` Python library to verify conformance with USPS Publication 28 addressing standards and OpenStreetMap conventions for geocoding.^{21,24} County names were assigned via spatial join to U.S. county TIGER/Line shapefiles and cross-validated against a ZIP code-to-county crosswalk file. A custom flag function was applied to identify parsing failures, including PO boxes and addresses missing a street number, which were excluded from the final dataset.

Reference locations were selected based on verified postal address quality and known GIS coordinates, enabling direct comparison between geocoder-produced coordinates and ground truth positions. As a quality check, each address's reference geographic coordinate was verified to fall within the same state listed in the address using a point-in-polygon spatial join. Of

the 432K locations, 272,963 input addresses with reference coordinates were geocoded, 2,395 failed the point-in-polygon check and were excluded. These failures did not reflect systematically poor reference data quality but rather appeared to stem from localized data transformation errors. After removing these records, 270,568 addresses remained as the final reference dataset.²⁵

Outlier Exclusion

Table 3. Summary of principled outlier removal method.

Records removed	Records remaining	Median	Mean	Max
2,438	270,445	82m	3,092m	1,850,246m

To reduce the influence of extreme positional errors on downstream accuracy calculations, one outlier removal approach was applied (Table 3). Principled outlier removal excluded records where the geocoded coordinate fell outside the expected state boundary based on a spatial join with 2025 Census TIGER/Line state boundary files, removing 2,438 records (0.89%). Separately, a spatial join on the reference ground truth coordinates identified 5 records where the reference coordinate itself fell outside the expected state boundary. These misannotations were also excluded as part of the 2,438 total removed, and are distinguished from the 2,433 geocoder misclassifications in the analysis that follows. This approach is conceptually grounded in Ratcliffe's (2001) use of cadastral records as an independent spatial reference to evaluate geocoding accuracy, and extends it to a national scale by using state boundaries as the validation layer.²⁶ Unlike percentile-based methods, the principled approach distinguishes between misclassification, where the geocoder placed a coordinate in the wrong state, and

misannotation, where the reference ground truth coordinate itself may be erroneous. Of the 2,438 flagged records, 2,433 were identified as misclassifications and only 5 as misannotations, indicating that state boundary mismatches in this dataset are attributable almost entirely to geocoder error.

Rural, Native Area, and Territory Classification

Reference locations were classified into four geographic categories using a single location classification variable (`loc_class`) to evaluate geocoding performance across area types that are typically underrepresented in research. Urban and rural designations were assigned using the USDA Rural-Urban Commuting Area (RUCA) codes, which classify census tracts based on population density, urbanization, and commuting patterns.²⁷ RUCA codes are based on commuting flows rather than strict land use, meaning that areas which appear geographically rural may receive urban designations if their residents primarily commute to nearby metropolitan centers, and vice versa.²⁷ Locations with RUCA codes of 1, 2, or 3 were classified as urban; all other RUCA codes were classified as nonurban (rural). American Indian, Alaska Native, and Native Hawaiian (AIAN/NH) area designations were assigned using tribal boundary maps developed by the Federal Communications Commission (FCC) to determine eligibility for unassigned 2.5 GHz wireless spectrum licenses.²⁸ These maps were used in place of Census TIGER/Line AIAN shapefiles, which have known coverage gaps for certain states.

The four resulting `loc_class` categories are defined as follows. A location is classified as "territory" if it falls within a U.S. territory, as territories do not have designated tribal areas and not all received RUCA designations. A location is classified as "tribal" if it falls within an AIAN/Native Hawaiian area regardless of urbanicity. A location is classified as "urban nontribal"

if it has a RUCA code of 1, 2, or 3 and falls within a U.S. state or the District of Columbia outside of a tribal area. A location is classified as "nonurban nontribal" if it has a RUCA code other than 1, 2, or 3 and falls within a U.S. state or the District of Columbia outside of a tribal area.

RESULTS

After outlier exclusion, geographic stratification showed substantial variation in geocoding accuracy across location types. Urban nontribal areas had the lowest median positional error at 76.4m (IQR 30.6–194.2m), followed by rural nontribal areas at 105.3m (IQR 45.6–306.9m) and rural tribal areas at 146.0m (IQR 62.8m–719.3m). U.S. territories showed dramatically higher median error at 4,838.5m (IQR 618.1m–13,924.3m), more than 60 times greater than urban areas. The wider interquartile ranges for rural and territory categories indicate not only higher median error but also greater variability in geocoding outcomes within these strata, suggesting less consistent geocoder performance in these geographies (Table 4).

Table 4. Geographic stratification summary comparing urban vs rural vs tribal vs territory DeGAUSS geocoder performance following principled outlier removal.

	Sample size	Median geocoding error (25th and 75th percentile)
Rural		
<i>Tribal</i>	3,243	146.0m (62.8m, 719.3m)
<i>Non-tribal</i>	50,361	105.3m (45.6m, 306.9m)
Urban	216,706	76.4m (30.6m, 194.2m)
Territories	134	4,838.5m (618.1m, 13,924.3m)

Records where both the reference coordinate and geocoded coordinate fell within the expected state boundary (Table 5, upper left quadrant) form the primary analysis dataset and are the basis for all positional error calculations reported in this study. Records falling into the remaining three quadrants in Table 5 were excluded from downstream analysis, as their distance calculations may reflect either geocoder error or reference data quality issues rather than true positional displacement.

Table 5. Classification of geocoding outcomes by reference coordinate accuracy and geocoded coordinate accuracy, based on principled spatial validation against 2025 Census TIGER/Line state boundaries.

	Reference coordinate correct (in expected state)	Reference coordinate incorrect (misannotation)
Geocoded coordinate correct (in expected state)	270,445 (98.56%)	0 (0%)
Geocoded coordinate incorrect (misclassification)	2,433 (0.89%)	5 (<0.01%)

Among three states of differing error ranges, California showed the lowest median positional error at 44.8m, Washington showed a moderate median error of 64.2m, and Alabama showed the highest at 106.1m. These three states were selected to illustrate the range of geocoding performance observed across the contiguous United States, spanning from the best-performing to the worst-performing states in the dataset. The difference in median error being more than double between California and Alabama demonstrates that geographic location alone plays a substantial role in determining positional accuracy (Table 6).

Table 6. Positional errors of low-error state (California), a mid-range state (Washington), and a high-error state (Alabama) based on precision of range geocodes after principled outlier removal.

	Sample size	Median geocoding error (25th and 75th percentile)
California	24,993	44.8m (18.9m, 118.8m)
Washington	5,135	64.2m (30.9m, 123.8m)
Alabama	3,236	106.1m (45.1m, 214.7m)

***Note: Table 6 includes all range-precision geocodes regardless of match score, while Tables 7a–7d restrict records to those meeting each specified match-score threshold**

The choropleth map of median positional error by state based on precision of range/street reveals a geographic clustering pattern in geocoding performance (Figure 4). The highest-error states are concentrated in the South, with Mississippi (116.4m), Alabama (106.1m), and Georgia (101.4m) showing the greatest median errors (Table 7). The lowest-error states are found predominantly in the Northeast and along the West Coast, with California (41.3m), New York (50.4m), and Connecticut (53.1m) performing best. This regional pattern does not align neatly with urbanicity, suggesting that regional factors such as road network documentation quality, address standardization practices, and land parcel size contribute independently to geocoding performance (Figure 4).

Figure 4. Choropleth map of median positional error by state, showing geographic variation across the U.S. (Puerto Rico is excluded from the color scale) based on precision of range/street.

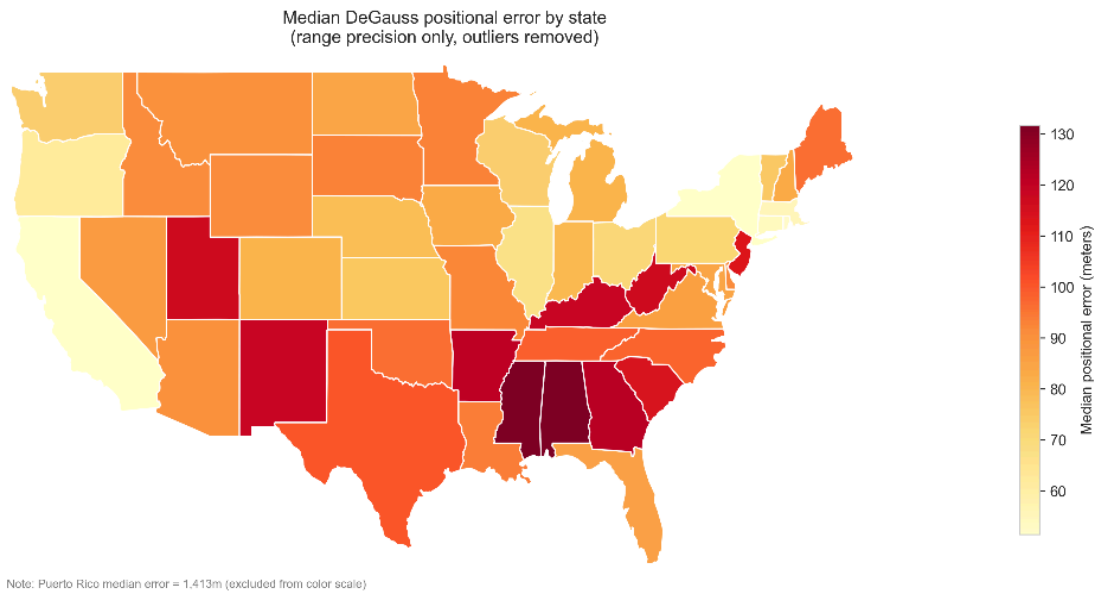


Table 7 presents state-level positional error for all states with at least 1,000 successful geocoding records at range or street precision and a match score of 0.8 or higher, sorted by median error in descending order. Mississippi (116.4m), Alabama (106.1m), and Georgia (101.4m) show the highest median errors, while New York (42.2m) and California (46.5m) show the lowest, with most states falling between 60m and 100m (Table 7).

Table 7. Geocoding positional error by state among states with $\geq 1,000$ successful geocoding records (precision of range or street, score ≥ 0.8 , principled outlier removal). Error was measured as straight-line distance in meters between geocoded and reference coordinates, and states were sorted by median error, descending.

State name	Count	Median error	Mean error	75th percentile	95th percentile
Mississippi	2057	116.4	1019.3	277.5	2716.8
Alabama	3336	106.1	644.7	235.3	1881.6
Georgia	8512	101.4	479.6	238.5	1463.2
South Carolina	4346	100.4	518.2	216.0	1522.3
Arkansas	2416	99.5	1124.2	217.0	2563.5
Kentucky	3381	98.2	870.0	214.8	2077.7

Utah	1401	97.6	7760.9	222.2	37593.0
New Mexico	1541	95.3	447.4	208.3	1346.8
Tennessee	6107	89.0	560.2	184.0	938.1
North Carolina	8330	88.0	468.9	175.2	894.7
Texas	20403	84.8	743.5	170.1	1145.2
South Dakota	1109	84.3	2484.2	201.3	2477.9
Oklahoma	3589	82.9	1035.3	177.4	1653.1
West Virginia	1653	82.0	432.7	224.3	1738.5
Louisiana	3746	81.7	785.1	174.4	1322.4
Arizona	4544	79.4	509.7	151.6	1162.2
Minnesota	4751	79.1	806.5	173.1	1449.9
Idaho	1477	78.0	3778.3	158.4	1208.2
Nevada	1656	77.3	11841.2	165.6	15150.9
Iowa	3206	77.2	862.5	145.0	752.4
Virginia	5847	76.8	351.6	150.0	597.1
Missouri	4967	76.7	775.6	164.9	1464.5
Montana	1062	76.3	2914.9	156.8	1127.5
Maryland	5929	75.8	241.7	148.2	474.7
Michigan	9306	72.7	377.5	151.0	696.7
Indiana	6622	72.2	535.5	152.2	902.5
Nebraska	2050	70.4	1618.0	132.2	1029.5
Florida	15502	69.3	393.4	155.0	981.9
Kansas	3035	69.2	838.8	127.8	781.1
Wisconsin	4288	68.7	423.6	134.6	678.9
Colorado	4162	66.5	360.7	139.8	873.0
Ohio	10776	66.4	381.7	145.3	866.4
Washington	5115	66.3	260.0	130.1	590.4
Pennsylvania	9324	64.8	315.3	150.8	593.2
Illinois	11106	61.4	473.5	123.1	817.5
Oregon	3095	54.7	2882.4	120.1	770.6
Connecticut	1179	50.6	301.4	124.1	552.2
Massachusetts	1875	49.2	399.4	130.4	604.1
California	25270	46.5	13172.5	131.5	82569.5
New York	14505	42.2	240.3	103.8	470.1

The boxplots confirm that positional error distributions are right-skewed across all three states, consistent with the log-normal distribution characteristic of street geocoding error noted by Zandbergen (2011). California shows the most compact interquartile range with a median of 51m, while Washington and Alabama show progressively wider distributions with medians of

74m and 132m respectively. All three states exhibit long upper tails extending to 10^5 meters or beyond, indicating that a small proportion of geocodes produce extreme positional errors even in the best-performing states (Figure 5a). Figure 5b breaks down each state's distribution further by match score threshold (100%, 90%, and 80%), with jittered points shaded and shaped by precision level (range, street, or zip) to show how precision composition shifts as the threshold is relaxed.

Figure 5a. Boxplots illustrating the within-state positional error variation and the contrast between a low-error state (California), a mid-range state (Washington), and a high-error state (Alabama). Whiskers extend to $1.5 \times \text{IQR}$ beyond the 25th and 75th percentiles.

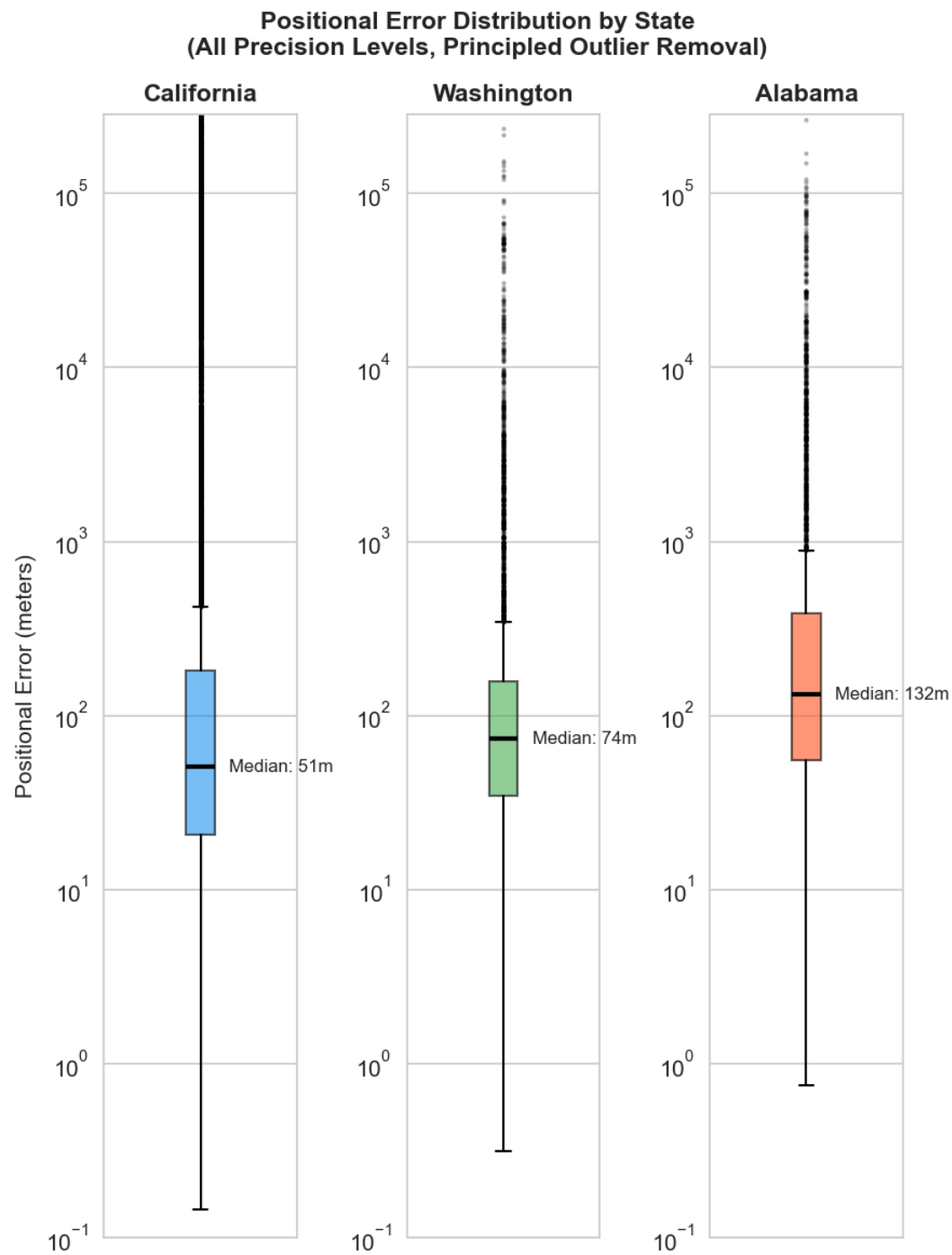


Figure 5b. Boxplots illustrate positional error by match score threshold for California, Washington, and Alabama.

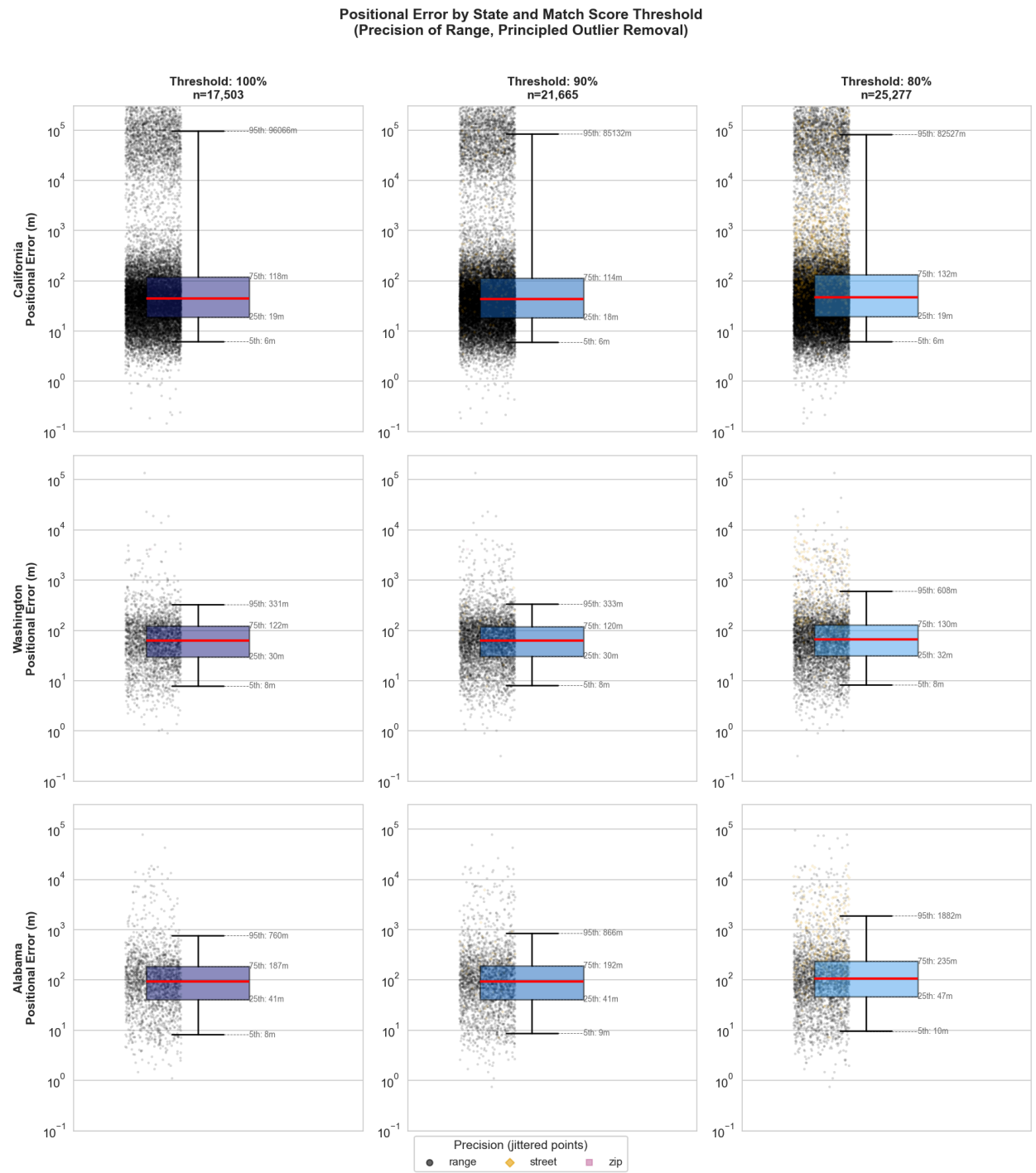
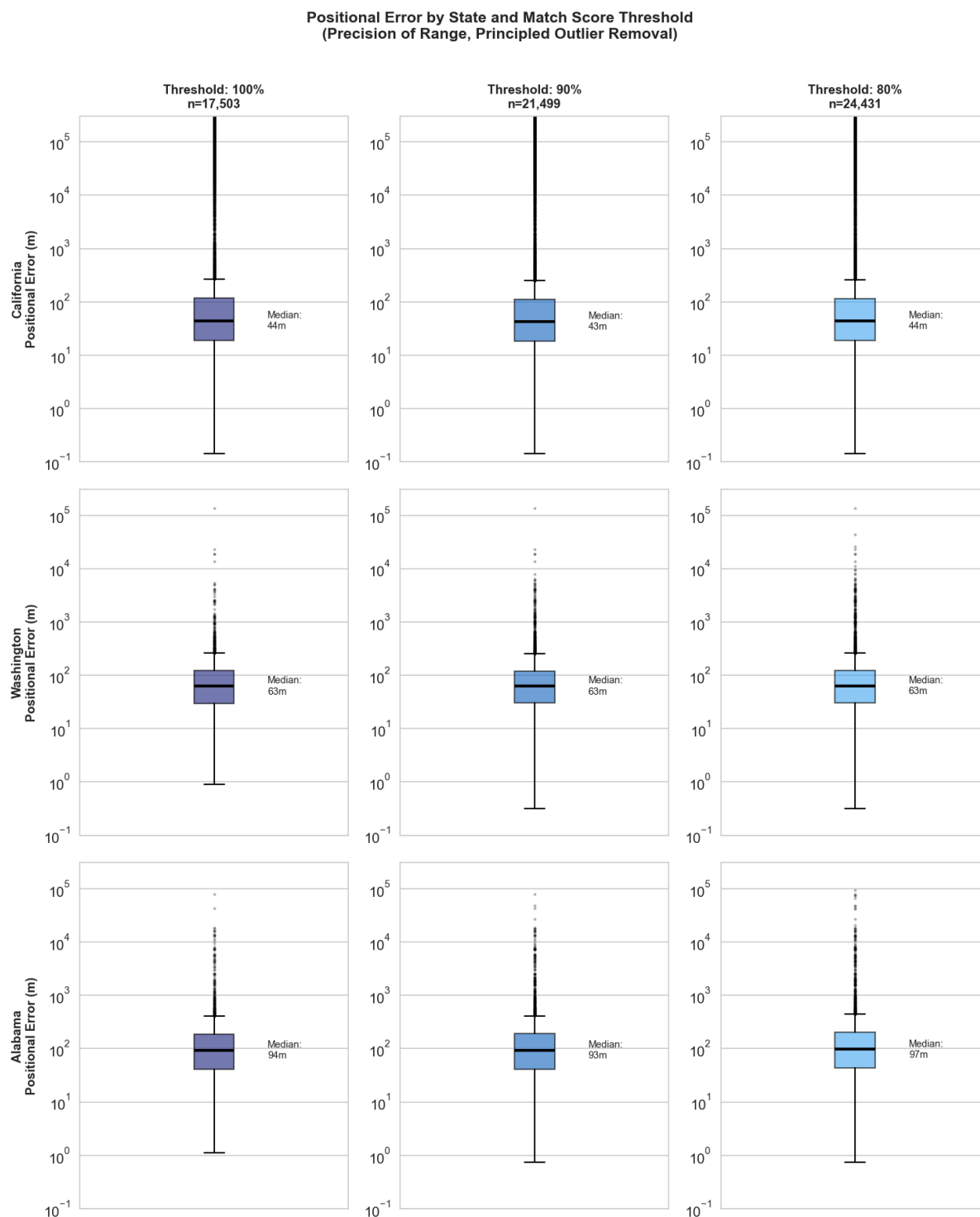


Figure 6 presents positional error distributions for California, Washington, and Alabama at match score thresholds of 100%, 90%, and 80%, restricted to range-precision geocodes following principled outlier removal. Across all three states, median positional errors were stable as the threshold decreased, consistent with the threshold sensitivity results reported in Tables 7b–7d. California showed median errors of 44m, 43m, and 44m at the 100%, 90%, and 80% thresholds respectively, while Washington remained at 63m across all three. Alabama’s median error showed a modest shift from 94m at 100% to 93m at 90% and 97m at 80%. The overall shape of each state's distribution, including the interquartile range and the extent of the upper tail, was consistent across thresholds, showing that relaxing the match score threshold from 100% to 80% does not substantially alter the distribution of positional error for range-precision geocodes within a given state (Figure 6).

Figure 6. Positional error distributions for three representative states at match score thresholds of 100%, 90%, and 80% (precision of range only, principled outlier removal). All panels share a common log-scale y-axis. Whiskers extend to $1.5 \times \text{IQR}$ beyond the 25th and 75th percentiles.



The threshold sensitivity analysis shows that relaxing the match score threshold from 100% to 60% substantially increases the amount of successfully geocoded records while producing only modest increases in median positional error across all states. Nationally, record count grows from 157,538 (58.3%) to 234,927 (86.9%), while median error rises only from 65.4m to 68.0m (Table 8a). However, the pattern varies by state. In California, median error remains nearly stable across all thresholds (44.4m at 100% to 44.8m at 60%), suggesting that lower-scoring records in California are not substantially worse in positional accuracy than higher-scoring ones (Table 8b). Washington shows a similar pattern of stability, with median error changing by only 1.4m across the full threshold range (Table 8c). Alabama, by contrast, shows a more pronounced increase where median error rises from 93.6m at 100% to 100.4m at 60%, alongside a larger gain in record count from 2,099 to 3,229 (Table 8d). This state-level comparison suggests that tradeoff between successfully geocoded records and accuracy is more consequential in higher-error states, where the additional records admitted at lenient thresholds tend to have meaningfully worse positional accuracy than those admitted at strict thresholds.

Table 8a. Threshold sensitivity summary for all states (precision of range only).

Threshold	Record count (% of total 270,445)	Median error (25th, 75th percentile)
100%	157,538 (58.3%)	65.3m (27.5m, 137.5m)
90%	204,427 (75.6%)	64.9m (27.4m, 136.6m)
80%	227,157 (84.0%)	66.9m (28.2m, 143.3m)
70%	232,131 (85.8%)	67.7m (28.4m, 145.9m)
60%	234,927 (86.9%)	68.0m (28.5m, 146.9m)

Table 8b. Threshold sensitivity summary for California (precision of range only).

Threshold	Record count	Median error (25th, 75th percentile)
100%	17,503	44.4m (18.8m, 118.4m)
90%	21,499	43.4m (18.4m, 112.5m)
80%	24,431	44.4m (18.7m, 116.5m)
70%	24,816	44.7m (18.8m, 118.0m)
60%	24,973	44.8m (18.8m, 118.5m)

Table 8c. Threshold sensitivity summary for Washington (precision of range only).

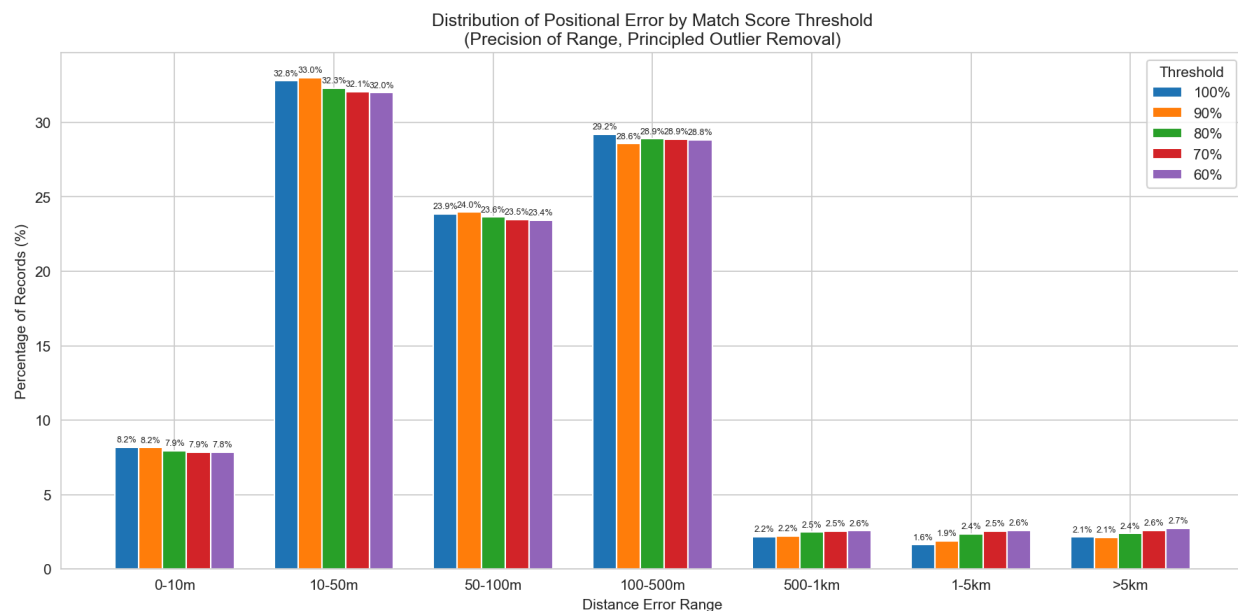
Threshold	Record count	Median error (25th, 75th percentile)
100%	2,815	62.6m (29.9m, 121.9m)
90%	4,447	62.5m (30.4m, 119.4m)
80%	4,883	63.2m (30.4m, 121.9m)
70%	4,993	63.9m (30.8m, 123.1m)
60%	5,118	64.0m (30.9m, 123.4m)

Table 8d. Threshold sensitivity summary for Alabama (precision of range only).

Threshold	Record count	Median error (25th, 75th percentile)
100%	2,099	93.6m (40.6m, 186.6m)
90%	2,733	93.1m (40.7m, 190.6m)
80%	3,074	97.3m (44.1m, 203.4m)
70%	3,176	99.3m (44.8m, 209.3m)
60%	3,229	100.4m (44.9m, 214.0m)

Figure 7 shows the percentage of precision of range geocodes falling into each positional error bin (0–10m, 10–50m, 50–100m, 100–500m, 500m–1km, 1–5km, and >5km) across the five match score thresholds, after principled outlier removal, for U.S. states. The distribution across error bins were mostly stable as the threshold decreased: about 8% of records fell within 0–10m, 32–33% within 10–50m, 24% within 50–100m, and 29% within 100–500m across all five thresholds. In contrast, the proportion of records in the highest-error bins increased modestly with more lenient thresholds, increasing from about 2.2% to 2.6% for the 500m–1km bin, 1.6% to 2.6% for the 1–5km bin, and 2.1% to 2.7% for the >5km bin. This pattern shows that relaxing the match score threshold disproportionately includes more high-error records, even though these records make up a small proportion of the overall dataset (Figure 7).

Figure 7. Percentage of range-precision geocodes by positional error bin and match score threshold, after principled outlier removal (U.S. states only).



The vast majority of valid geocodes (87%) were assigned precision of range, the highest available precision level in DeGAUSS, indicating that the geocoder successfully matched most

records to a specific address range along a street segment. An additional 12% were matched at street precision. Fewer than 1% of records were assigned zip, intersection, or city precision (Table 9).

Table 9. DeGAUSS precision level distribution for valid geocodes after principled outlier removal (n=270,445).

Precision Level	Count (% of total)
Range	235,282 (87%)
Street	32,848 (12%)
Zip	2,207 (0.82%)
Intersection	105 (0.04%)
City	3 (<0.01%)

Territory-level results reveal geocoding performance that is categorically different from any U.S. state. Guam showed the highest median error at 4,249m (IQR 2,341–5,424m), followed by Northern Mariana Islands at 2,319m, U.S. Virgin Islands at 1,967m (IQR 1,640–2,294m), and Puerto Rico at 450m (IQR 127.9–1,596m). Puerto Rico had a median error more than ten times greater than California. The extreme error magnitudes across all four territories suggest that standard DeGAUSS geocoding using TIGER/Line reference data is not adequate for reliable geomarker linkage in U.S. territory populations (Table 10).

Table 10. Individual territory results with sample size, median error, and IQR.

Territory	Sample size	Median Error (25th, 75th percentile)
Puerto Rico	71	450.0m (127.9m, 1,596m)

Guam	58	4,249m (2,341m, 5,424m)
U.S. Virgin Islands	3	1,967m (1,640m, 2,294m)
Northern Mariana Islands	2	2,319m (2,319m, 2,319m)

DISCUSSION

This study found that relaxing the DeGAUSS match score threshold from 100% to 60% substantially increased the amount of successfully geocoded records, from 58.3% to 86.9% of the dataset, while producing only modest increases in median positional error nationally. However this tradeoff between the amount of successfully geocoded records and accuracy was more consequential in higher error states, and geographic stratification revealed substantially higher positional error in U.S. territories, more than that in rural and tribal areas. A principled spatial outlier removal method, which excluded records where the geocoded coordinate fell outside the expected state boundary, further distinguished geocoder misclassification from reference coordinate error, showing that the majority of flagged records reflected geocoder error rather than reference data error. The implications of these findings for geocoding quality assessment and precision medicine research are discussed below.

Geocoding Quality and the Choice of Success Threshold

A central finding of this work is that no single match score threshold uniformly captures high-quality geocodes across all U.S. geographies. When the match score threshold is relaxed from 100% to 60%, the number of records classified as successfully geocoded increases substantially from 157,538 (58.3% of the dataset) to 234,927 (86.9%) (Table 8a). However, the effect on positional error is concentrated in the upper tail rather than uniformly distributed. The

median positional error showed modest sensitivity to this relaxation, rising from 65.3m to 68.0m. The 95th percentile increased more substantially, from 430m to 572m, indicating that the primary consequence of relaxing the threshold is an enlargement of the high-error tail rather than a uniform degradation of accuracy across all records. These results suggest that applying a single threshold across a geographically diverse national dataset risks either excluding a disproportionate share of records at strict thresholds or including a higher volume of high-error geocodes at lenient ones, with the tradeoff being most consequential for analyses sensitive to extreme positional displacements.

It is also notable that even at the strictest 100% threshold, approximately 23% of records had positional errors greater than 500m, including 9.7% between 1 and 5km and 7.9% exceeding 5km (Figure 7). This indicates that threshold alone is insufficient to guarantee positional accuracy, and that geographic stratification and outlier removal are necessary complementary steps in any quality assessment framework. This observation aligns with Zandbergen (2011), who noted that the distribution of positional error in street geocoding more closely resembles a log-normal distribution than a normal one, making the mean a poor summary statistic and reinforcing the use of median and percentile-based measures to characterize the full error distribution.²⁹

Brokamp et al. (2018) made a related observation in their validation of DeGAUSS against ArcGIS, Google, and SAS using residential addresses in Hamilton County, Ohio, noting that the choice between increased address inclusion and decreased geocoding accuracy represents a fundamental tradeoff between selection bias and exposure assessment bias.¹⁶ Our findings extend this observation to a national scale and across heterogeneous geographic settings, showing that the magnitude of this tradeoff varies by geography.

Defining Success: How Match Criteria Compare Across Studies

A key methodological challenge in geocoding quality assessments is defining what constitutes a successful geocode. Prior studies have approached this differently, and those differences bear directly on how results can be compared. Zhan et al. (2006) required both a specific match code indicating exact USPS address database matches or intersection-level matches and a location code, ensuring the geocoded point was placed at the house range level with the highest available positional accuracy. ArcGIS geocoding in the same study used a minimum match score of 70 out of 100, with spelling sensitivity set to 50.³⁰ Zandbergen (2011) used a minimum match score of 80 in ArcGIS 9.1 across three reference datasets, distinguishing perfect matches (score of 100) from less-than-perfect matches (score between 80 and 100) and flagging ties separately as a quality concern.²⁹ These threshold choices are not equivalent and do not match DeGAUSS's scoring system, which produces a continuous similarity score on a 0–1 scale reflecting textual agreement between the input address and the matched reference address, interpreted within precision levels rather than across them. The present study evaluates five thresholds from 60% through 100% applied to range-precision geocodes only, which corresponds most closely to the "house range" or "street range" success criteria used in prior work.

Precision Level and Geomarker Linkage

The precision level assigned by DeGAUSS, ranging from range interpolation at the highest end to ZIP code or city centroid at the lowest, carries implications for the reliability of downstream geomarker linkages. Brokamp et al. defined successful geocoding as precision of "range" or "street" and excluded ZIP and city-level results from geomarker assessment, a

threshold our findings support as a reasonable starting point.¹⁶ In the present dataset, 87% of records were geocoded at address range-level precision and 12% at street-level precision, with fewer than 1% at zip, intersection, or city level, suggesting that the vast majority of records meet this quality bar before any score-based filtering is applied (Table 8).

The consequences of precision level vary by the type of geomarker being linked. Area-based geomarkers such as census tract median income, neighborhood deprivation indices, or county-level disease rates are more tolerant of moderate positional error because spatial correlation within geographic units buffers small displacements.³¹ Brokamp et al. found that even with a median geocoding error of 40 meters, DeGAUSS assigned 95.5% of addresses to the correct census tract and achieved a Pearson correlation of 0.97 for census tract median household income compared to the gold standard.¹⁶ The median errors observed in this study for U.S. states, ranging from 44.8m in California to 116.4m in Mississippi, are broadly consistent with the range of 25–168m reported across prior studies of residential address geocoding, as summarized by Zandbergen (2011).²⁹ By contrast, geomarkers that depend on exact location such as distance to the nearest roadway proximity to a pollution source, or distance to the nearest healthcare facility are much more sensitive to positional error. However, the practical impact of that error depends on the scale of analysis. Zhan et al. (2006) found that despite significantly different positional accuracy between two geocoding methods (median errors of 69m versus 48m), sensitivity and specificity for classifying addresses within 1,500m of toxic release inventory facilities were nearly identical.³⁰ This suggests that moderate geocoding error has limited practical impact when buffer distances are large, and this finding should not be generalized to fine-scale analyses where buffer distances are small or where the spatial distribution of exposures changes rapidly across short distances.

Geographic Variation in Geocoding Performance

Among U.S. states, the highest median positional errors were observed in Mississippi (116.4m), Alabama (106.1m), and Georgia (101.4m), while Massachusetts (49.2m), California (46.5m), and New York (42.2m) showed the lowest (Table 7). This variation does not follow a simple urban-rural pattern as Mississippi, Alabama, and Georgia are states with majority rural geographic area, suggesting that regional factors beyond urbanicity, such as road network density, address standardization practices, and the currency of the underlying TIGER/Line reference data, contribute to geocoding performance differences across states (Figure 4). Zandbergen (2011) found that rural single-family residential addresses consistently showed lower positional accuracy than urban single-family residential addresses, attributing this to longer road segments, larger parcels, and greater parcel size variability that limit the accuracy of linear interpolation along a street segment.²⁹ This mechanism likely contributes to the higher errors observed in states with larger rural footprints and less densely documented road networks. The Southern states with highest error in our dataset are not exclusively rural, indicating that urbanicity alone does not fully explain the observed heterogeneity.

U.S. territories showed dramatically higher positional error than any U.S. state. Guam had a median error of 4,248.9m, Northern Mariana Islands 2,319m, U.S. Virgin Islands 1,967m, and Puerto Rico 450.0m (Table 10). It should be noted that the territory sample was small (n=134 records after outlier removal), and some portion of this error may reflect limitations in the public reference datasets used rather than systematic geocoder underperformance. Territory-level errors of this magnitude make standard address-range geocoding unsuitable for reliable geomarker linkage in these populations, and alternative approaches should be considered.

Implications for Environmental and Precision Medicine Research

Geocoding is a useful tool in public health research and practice that can connect health data to geographic locations, help identify vulnerable populations, and reveal common characteristics associated with risk to inform targeted interventions. Geocoding also supports health equity and accessibility, helping to reveal disparities in access to care, healthy food, green space, and other resources across communities. Within cities, precise geocoding can aid in detection of clusters of infectious disease transmission or poor air quality exposure events, which is essential for rapid response and resource allocation. Over time, spatial health data can help syndromic surveillance identify environmental, social, or infrastructural factors that drive poor health outcomes.^{32,33}

The All of Us Research Program, which motivated the operational geocoding work that preceded this benchmarking study, aims to link participants' residential histories to environmental, social, and structural determinants of health at a national scale.¹⁸ Achieving this goal reliably depends on geocoding quality that is both accurate and well-characterized. Inaccurate geocodes introduce exposure misclassification; assigning an individual to the wrong neighborhood, census tract, environmental exposure zone, misconstruing work address as the residential address, and misinterpreting the temporal order and time boundaries at each residence across multiple addresses can bias associations between place-based exposures and health outcomes.

The implications are especially prominent for precision medicine applications. Precision medicine initiatives seek to understand how individual-level variation in genetics, environment, and lifestyle interact to shape health. Geographic location is a critical input into this framework,

as it determines ambient exposures to air pollutants, access to care, neighborhood socioeconomic conditions, and built environment characteristics. While ambient exposures within a block are often approximately the same, geocoding errors large enough to displace an individual across neighborhoods, census tracts, or environmental exposure zones can introduce substantial misclassification. Where geocoding quality is systematically lower as seen in U.S. territories and in high-error states, environmental data linked to those individuals will be less accurate.

Infectious Disease Surveillance and Public Health Applications

Beyond precision medicine, geocoding quality has direct consequences for infectious disease surveillance and broader public health practice. Hotspot detection methods, which identify geographic clusters of elevated disease incidence, depend on the spatial accuracy of case addresses to correctly assign individuals to the appropriate location.³⁴ In U.S. territories specifically, where median positional error exceeded 450m even in Puerto Rico and reached over 4,000m in Guam, errors of this magnitude would be sufficient to misplace a case across neighborhood or even municipal boundaries.

Studies examining healthcare accessibility similarly depend on accurate geocoding. Positional displacements of several hundred meters to several kilometers, as observed in territories and in the high-error tail across states, can shift an individual's classification between adequate and inadequate access depending on the spatial distribution of nearby facilities.

Distinguishing Geocoder Error from Reference Data Error

One methodological contribution of this study is the distinction between misclassification, where the geocoder produces a coordinate outside the expected state boundary,

and misannotation, where the reference ground truth coordinate itself is incorrect. Critically, the analysis revealed a near-complete asymmetry: Of the 2,438 records (0.89%) excluded by the principled method, 2,433 were geocoder misclassifications (coordinates placed in the wrong state) while only 5 were misannotations, meaning the reference ground truth coordinate itself fell outside the expected state boundary. The most common state-level confusions involved geographically adjacent states (e.g. California and Utah, Oregon and California, Utah and California), consistent with known limitations of range-based interpolation geocoders near state borders. For territories, where the small sample and dramatically higher errors make the misclassification-misannotation distinction harder to resolve, further investigation with curated reference data is needed. Future benchmarking studies using public reference datasets should consider this distinction, as reference coordinate quality may vary across source datasets and geographic regions in ways that are not always apparent from the data alone.

CONCLUSION

This study evaluated the performance of the DeGAUSS geocoder using a curated reference dataset of over 270,000 public locations spanning U.S. states and territories, with a focus on how the choice of quality threshold affects geocoding outcomes and how positional error varies across geographic strata. First, the sensitivity analysis demonstrates that the relationship between match score threshold and geocoding quality is not straightforward. Relaxing the threshold from 100% to 60% substantially increases the number of records classified as successfully geocoded while producing only slight increases in median positional error. For most area-based geomarker applications, a threshold in the range of 60% to 80% may balance the amount of successfully geocoded records and accuracy well. For applications

sensitive to exact location like proximity-based exposure assessments or distance-to-facility analyses, stricter thresholds or additional spatial validation steps are needed.

Second, geographic stratification reveals that a single threshold is not equally appropriate across all U.S. geographies. Median positional error ranged from 41.3m in California to 116.4m in Mississippi among states, with the highest-error states concentrated in the South rather than following a simple urban-rural gradient. Rural tribal areas showed higher median errors (146.0m) than rural non-tribal areas (105.3m) or urban areas (76.4m), consistent with known limitations of range-based interpolation in areas with less complete road network documentation.²⁹ U.S. territories showed dramatically higher positional errors than any U.S. state, suggesting that standard DeGAUSS geocoding is insufficient for reliable geomarker linkage in these populations and that territory-specific approaches are needed.

Third, the principled outlier removal method excluding records where the geocoded coordinate falls outside the expected state boundary provides a spatially grounded alternative to percentile-based outlier removal. Of the 2,438 records flagged by the principled method, 2,433 were geocoder misclassifications and only 5 were reference data errors, confirming that state boundary mismatches in this dataset are driven almost entirely by geocoder error rather than poor reference coordinate quality.

This study's strength is that the reference dataset spans over 270,000 locations across all U.S. states and territories, offering a level of geographic diversity not captured in previous geocoding benchmarking studies, which have typically been limited to single cohorts or geographically bounded regions.

This study also has limitations. The reference dataset consists of public locations such as hospitals, schools, banks, and places of worship, which can differ in address format,

documentation quality, and underlying road network accuracy from residential addresses. As a result, the positional error estimates reported here may not generalize directly to residential geocoding performance. Moreover, excluding PO Box addresses and addresses missing a street number during reference data curation may disproportionately affect rural, tribal, and territory locations, where postal delivery infrastructure differs from urban norms (e.g., PO Box-only delivery routes, unnumbered addresses on named family roads). This could bias the reference dataset toward more standardized addresses within these strata, potentially underestimating true geocoding error in the populations already shown to have the highest positional error in this study. Additionally, the territory sample ($n=134$ after outlier removal) is small, especially for the U.S. Virgin Islands ($n=3$) and Northern Mariana Islands ($n=2$), limiting the precision of territory-specific error estimates. Finally, this study evaluated a single geocoding framework DeGAUSS using a single version of TIGER/Line reference data. This means that geocoding performance may differ for other geocoding systems or reference sets, so the findings should not be assumed to generalize beyond the specific software and reference data evaluated in this work.

Together, these findings inform the following guidelines for researchers using DeGAUSS or similar interpolation-based geocoders on large, geographically diverse U.S. datasets: report match score thresholds and outlier removal methods explicitly; analyze positional error separately for territories and rural areas rather than aggregating across all geographies; use median and percentile-based statistics rather than the mean to summarize error distributions; and consider the type of geomarker being linked when deciding how much positional error is within acceptable bounds for the analysis. As the All of Us Research Program and similar national precision medicine initiatives continue to expand, establishing principled, geography-aware

standards for geocoding quality assessment will be essential for ensuring that residential linkage is accurate and equitable across the full diversity of the U.S. population.

REFERENCES

1. Mitchell R, Popham F. Effect of exposure to natural environment on health inequalities: an observational population study. *The Lancet*. 2008;372(9650):1655-1660. doi:10.1016/S0140-6736(08)61689-X
2. Balbus JM, Malina C. Identifying vulnerable subpopulations for climate change health effects in the United States. *J Occup Environ Med*. 2009;51(1):33-37. doi:10.1097/JOM.0b013e318193e12e
3. Dwyer-Lindgren L, Bertozzi-Villa A, Stubbs RW, et al. Inequalities in Life Expectancy Among US Counties, 1980 to 2014: Temporal Trends and Key Drivers. *JAMA Intern Med*. 2017;177(7):1003-1011. doi:10.1001/jamainternmed.2017.0918
4. What is geocoding?—ArcMap | Documentation. Accessed October 11, 2025. <https://desktop.arcgis.com/en/arcmap/latest/manage-data/geocoding/what-is-geocoding.htm>
5. Mazumdar S, Rushton G, Smith BJ, Zimmerman DL, Donham KJ. Geocoding accuracy and the recovery of relationships between environmental exposures and health. *Int J Health Geogr*. 2008;7:13. doi:10.1186/1476-072X-7-13
6. Faure E, Danjou AMN, Clavel-Chapelon F, Boutron-Ruault MC, Dossus L, Fervers B. Accuracy of two geocoding methods for geographic information system-based exposure assessment in epidemiological studies. *Environ Health Glob Access Sci Source*. 2017;16(1):15. doi:10.1186/s12940-017-0217-5
7. Shah TI, Bell S, Wilson K. Geocoding for public health research: Empirical comparison of two geocoding services applied to Canadian cities. *Can Geogr Géographe Can*. 2014;58(4):400-417. doi:10.1111/cag.12091
8. Zandbergen PA. Geocoding Quality and Implications for Spatial Analysis. *Geogr Compass*. 2009;3(2):647-680. doi:10.1111/j.1749-8198.2008.00205.x
9. Beck AF, Sandel MT, Ryan PH, Kahn RS. Mapping Neighborhood Health Geomarkers To Clinical Care Decisions To Promote Equity In Child Health. *Health Aff (Millwood)*. 2017;36(6):999-1005. doi:10.1377/hlthaff.2016.1425
10. Beck AF, Anderson KL, Rich K, et al. Cooling The Hot Spots Where Child Hospitalization Rates Are High: A Neighborhood Approach To Population Health. *Health Aff (Millwood)*. 2019;38(9):1433-1441. doi:10.1377/hlthaff.2018.05496

11. Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. *Am J Public Health*. 2001;91(7):1114-1116. doi:10.2105/ajph.91.7.1114
12. Kinnee EJ, Tripathy S, Schinasi L, et al. Geocoding Error, Spatial Uncertainty, and Implications for Exposure Assessment and Environmental Epidemiology. *Int J Environ Res Public Health*. 2020;17(16):5845. doi:10.3390/ijerph17165845
13. Swift J, Wilson J. *Geocoding Best Practices: Reference Data, Input Data and Feature Matching*. 2008.
14. Goldberg DW, Ballard M, Boyd JH, et al. An evaluation framework for comparing geocoding systems. *Int J Health Geogr*. 2013;12:50. doi:10.1186/1476-072X-12-50
15. Han D, Bonner MR, Nie J, Freudenheim JL. Assessing bias associated with geocoding of historical residence in epidemiology research. *Geospatial Health*. 2013;7(2):369-374. doi:10.4081/gh.2013.93
16. Brokamp C, Wolfe C, Lingren T, Harley J, Ryan P. Decentralized and reproducible geocoding and characterization of community and environmental exposures for multisite studies. *J Am Med Inform Assoc JAMIA*. 2018;25(3):309-314. doi:10.1093/jamia/ocx128
17. Vancil A, Manning E, Brokamp C. *Software Application Profile: DeGAUSS – Decentralized Geomarker Assessment for Multi-Site Studies*. 2023. doi:10.20944/preprints202303.0008.v1
18. All of Us Research Program Overview. Accessed May 31, 2026. <https://allofus.nih.gov/article/program-overview>
19. Youngbloom AJ, Rowland M, Mooney SJ, Singer K, Szpiro A, Saelens BE. Investigating the Role of the Environment on Physical Activity Interventions (the InSPACE Project): Protocol for a Pooled Secondary Analysis of Randomized Controlled Trials. *JMIR Res Protoc*. 2026;15:e83151. doi:10.2196/83151
20. Homeland Infrastructure Foundation-Level Data (HIFLD) | Homeland Security. Accessed June 4, 2026. <https://www.dhs.gov/gmo/hifld>
21. Publication 28 - Postal Addressing Standards | Postal Explorer. Accessed March 2, 2026. <https://pe.usps.com/text/pub28/welcome.htm>
22. Brokamp C. Releases · degauss-org/geocoder. GitHub. Published online March 2, 2026. Accessed June 11, 2026. <https://github.com/degauss-org/geocoder/releases>
23. What is a Container? | Docker. Accessed March 13, 2026. <https://www.docker.com/resources/what-container/>
24. Gregg F, Deng C, Batchkarov M, Cochrane J. datamade/usaddress. GitHub. Published online May 26, 2026. Accessed June 4, 2026. <https://github.com/datamade/usaddress>

25. Chang BCY. brian-cy-chang/UW_Geospatial. GitHub. Published online December 23, 2024. Accessed June 4, 2026. https://github.com/brian-cy-chang/UW_Geospatial
26. Ratcliffe JH. On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units. *Int J Geogr Inf Sci*. 2001;15(5):473-485. doi:10.1080/13658810110047221
27. Rural-Urban Commuting Area Codes | Economic Research Service. Accessed June 4, 2026. <https://www-tx.ers.usda.gov/data-products/rural-urban-commuting-area-codes>
28. 2.5 GHz Rural Tribal Maps | Federal Communications Commission. Accessed June 4, 2026. <https://www.fcc.gov/25-ghz-rural-tribal-maps>
29. Influence of street reference data on geocoding quality: Geocarto International: Vol 26, No 1. Accessed May 31, 2026. <https://www.tandfonline.com/doi/abs/10.1080/10106049.2010.537374>
30. Zhan FB, Brender JD, De Lima I, Suarez L, Langlois PH. Match rate and positional accuracy of two geocoding methods for epidemiologic research. *Ann Epidemiol*. 2006;16(11):842-849. doi:10.1016/j.annepidem.2006.08.001
31. Dow D. Mapping Health Disparities: *Del J Public Health*. 2024;10(1):106-110. doi:10.32481/djph.2024.03.14
32. Karras B. Comparison of automated geocoding methods in a distributed multi-county disease surveillance system. Published online January 1, 2007. Accessed July 9, 2026. https://www.academia.edu/99423669/Comparison_of_automated_geocoding_methods_in_a_distributed_multi_county_disease_surveillance_system
33. Karras BT, Bliss D, Horn S, Metcalf H, Wagner S, Lober B, Lindquist S. Syndromic Surveillance Information Collection – Geocodes for Urban to Rural Mixed Environments: SSIC-Geo. *Adv Dis Surveill*. 2006;1:38. Accessed July 13, 2026. <https://knowledgerepository.syndromicsurveillance.org/syndromic-surveillance-information-collection-geocodes-urban-rural-mixed-environments-ssic-geo>
34. Lessler J, Azman AS, McKay HS, Moore SM. What is a Hotspot Anyway? *Am J Trop Med Hyg*. 2017;96(6):1270-1273. doi:10.4269/ajtmh.16-0427