TRACKING PROGRAM GEOCODING STANDARDS

January 2023

Please contact trackingsupport@cdc.gov with any questions.

Purpose

The Tracking Program has been working to increase the availability and accessibility of sub-county data. The Tracking Program's broader sub-county efforts include the following:

- Developing geocoding guidelines for assigning address-level health data to census tracts
- Using geographic aggregation to create standardized sub-county geographies to allow for data to be displayed at a finer scale than county-level data
- Evaluating current suppression rules used by data stewards for protecting confidentiality and developing new suppression rules for sub-county data
- Assessing population estimate datasets to determine how the use of different denominator datasets affect rates and their interpretation

This document addresses the first point and defines geocoding standards for the Tracking Program. These standards will allow the Tracking Program to present consistent data and supporting information to users of the Tracking Network. This document also serves as a reference for other programs to uniformly geocode information. The standards presented here maximize the balance between including addresses and meeting a reasonable standard of address quality.

This document only provides guidance for sub-county geocoding. For now, the Geospatial Workgroup has decided that county level standards will not be developed. This was due to decisions that were made within the workgroup, including geocoding to finer spatial resolution data (i.e., census tract) requiring more precise geocodes and more guidance on how to treat those records when they cannot be reliably placed within a geographic unit. The workgroup did, however, decide to include imputation best practices, which can also apply to county-level data.

Geocoding Standards

The Standards and Network Development (SND) Geospatial Team voted on a set of standards in July and August of 2018. The standards outline how geocoding of records should be handled, including special cases (e.g., rural routes and group quarters), and are described below.

1) The Tracking Program will only accept data that have been geocoded to the census tract level. Data may be displayed at the census tract level, aggregations of census tracts, county level, or state level, depending on data stability and confidentiality concerns.

2) All data for the years 2019 or earlier will be geocoded to census tract boundaries from Census 2010. After 2019, geocoding will move forward with each new decennial census decade, leaving previously geocoded data as is.

3) All geocoded records will be classified into the following categorization scheme (i.e., buckets):
   - High precision
   - Low precision
   - County precision
   - State precision
   - Unknown precision

   (*See the suggested process in Appendix A below.*)

4) The Tracking Program recommends against geocoding records with P.O. Box addresses directly to the census tract level (i.e., to the post office where the boxes are located) as this might inflate the number of cases for those census tracts. Instead, records with P.O. Box addresses should be separated out during the data cleaning phase of the geocoding process and either removed or imputed. In all cases, records with P.O. Box addresses should be classified as low precision.

   In dealing with records with P.O. Box addresses, the following point should be documented:

   > How recipients handle P.O. Box records. For example, P.O. Box records may be imputed to census tract from an alternative sub-county geography (e.g., ZIP code, town) following a recipient's imputation method of choice. Alternatively, a recipient could choose to remove P.O. Box records from sub-county submission if they include how many records were removed as part of the metadata.

5) Excluding P.O. Boxes, any record that geocodes to a geography that fits entirely within a single census tract (e.g., town or ZIP code) and meets the appropriate minimum match criteria for the geocoding software should be treated as high precision. In addition, excluding P.O. Boxes, an address that is matched to a geography level of ZIP+4 and meets the appropriate minimum match criteria for the geocoding software can also be classified as high precision. The minimum match criteria for ESRI, MapMarker/Centrus, and the Texas A&M Geocoder are defined below in the "How to Geocode" section.

6) Records with rural route addresses should be geocoded as normal using the geocoding process outlined below.

   Records with rural route addresses should be geocoded as follows:
   - If the rural route geocode has a road segment that falls completely within a tract, manual geocoding can place the rural route in high precision.
   - If the rural route geocode results in a ZIP code centroid, then the rural route must be placed in low precision.

7) Records associated with group quarters addresses, as defined by the Census Bureau (https://www.census.gov/topics/income-poverty/poverty/guidance/group-quarters.html), should be geocoded as normal using the geocoding process outlined below.

8) Recipients may propose a new minimum match criteria standard for additional geocoding software other than the software already considered by the SND Geospatial Team (i.e., ESRI, MapMarker/Centrus, Texas A&M Geocoder). The process is outlined as follows.

Follow these steps to define a new minimum match criteria for a different software:

1. Identify the software.
2. Identify what result value/code/identifier will be used to measure geocoding precision.
3. Communicate the cut points of the result value and how they will fit into the categorization scheme (high, low, unknown, county, state).
4. Identify any other variances from the standards that will occur by using this software and how those will be handled.
5. The SND workgroup will assign a team to review the proposal and work with the recipients to create an acceptable documentation of the proposal.
6. The documentation will be added as an appendix to the standards document.

Geocoding Process

Geocoding is an iterative process; steps may be repeated and don't necessarily need to be completed in this order.

1. Data Cleaning

- Identify and remove duplicate records. Refer to the how to guide for the particular dataset you are geocoding. Note that duplicate addresses may be valid and do not necessarily indicate duplicate records.
- Identify and separate P.O. Boxes addresses. Do not place all P.O. Boxes into the census tract of the post office. Check the standards above for more information.
- Match and replace common abbreviations/misspellings/synonyms for street names.
- Many batch geocoding solutions perform address standardization, auto correction of common address errors, and can help identify records in need of further attention in the data cleaning process. Alternatively, you can use address standardization software.

2. Geocoding Records

- Geocode the cleaned input records resulting from step 1 using the geocoding software of choice.
- Note any departure from default software settings; these will need to be discussed in the metadata.

3. Sorting Into Categorizations (Buckets)

After addresses have been geocoded, use the following categorization scheme to aggregate the individual records into the buckets defined below.

High precision: Any record that geocodes to a geography that fits entirely within a single census tract and meets the minimum match criteria defined for the software will be treated as high precision.

The minimum match criteria outlined by the current standards are as follows:

- For geocoding software that produces a match score as standard output (e.g., ESRI, Texas A&M Geocoder): match score **≥ 85**.
- For MapMarker/Centrus: follow instructions for location codes (*see Appendix B for code definitions*).

Low precision: Either 1) the record geocoded to a geography that fit entirely within a single census tract but did not meet the minimum match criteria for high precision OR 2) the record did not fit entirely within a single census tract but can be imputed to a census tract from an alternative sub-county unit (e.g., town, ZIP code).

County only: Either the input county is known or the county identified by the geocoding software is reliable. Report at county level.

If there is a discrepancy between county given by the geocoding software and county given in the dataset (which might be subject to typos/human error), the geocoding software county assignment should be prioritized.

State only: Address cannot be reasonably matched to a county or any sub-geography, but state of residence is known. Report at state level.

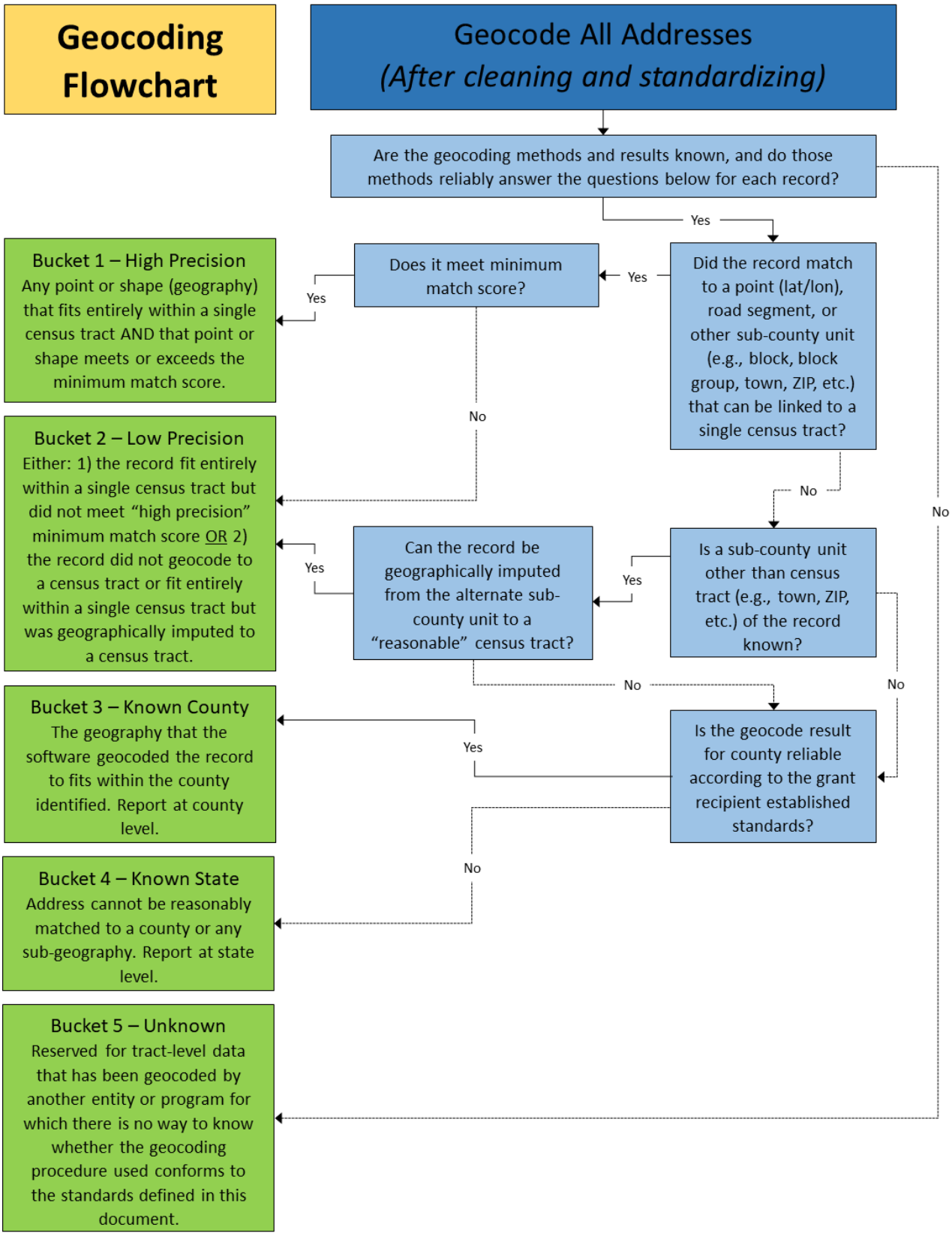
Unknown precision: Reserved for tract-level data that have been geocoded by another entity or program for which there is no way to know whether the geocoding procedure used conforms to the standards defined in this document.

4. Documentation of the Process

The metadata requirements for sub-county data are listed below. It is important for data users to have access to the information about the geocoding metadata to understand limitations of the data. If the Tracking Network Geocoding Standards were not used or if the data arrived geocoded by another entity or person, it might still be possible to use this information to assign geocoded data to the appropriate precision category. This might involve a discussion with the data steward about the process that was used. Questions to consider include the following:

1. Who geocoded the data?
2. What was the date the data were geocoded?
3. Were data geocoded using the Tracking Network Geocoding Standards?
4. What geocoding software was used? Include software name (e.g., ESRI, MapMarker/Centrus, Texas A&M Geocoder, etc.) and version. What type of address locator did you use (e.g., rooftop, street network, centroid, parcel)?
5. What underlying reference database(s) were used by the geocoding software (e.g., E911, TIGER, TomTom)?
6. Describe any changes made to default geocoding settings.
7. How were geocoded records matched to a census tract? That is, did the software used provide this information as standard output, or was some other method used for determining whether the geographic coordinates assigned to a record could be associated with a geography that fits entirely inside of a single census tract?
    a. Was imputation used to derive the census tract? If so, describe the method used.
    b. If imputation was not used, why not?
    c. What percentage of your data were you able to assign a tract to?
8. How many individual records were in the input dataset? What was the overall match percentage?
9. How many records with P.O. Box addresses were in the input dataset?
10. Were records with P.O. Box addresses removed or imputed to census tract? If imputed, describe the imputation process that was used.

Appendix A. Sub-County Specific Flowchart



Appendix B. MapMarker/Centrus Equivalent for Match Score

How to get MapMarker/Centrus equivalent of "≥ 85 Match Score."

High Precision

Location codes with:

1st character = A or 2nd character = B or T

A***, *B** or *T**

Low Precision

Location codes with:

4th character = W or 1st three characters = ZC9

***W, ZC9*

Appendix C. Block to Tract Resources

Block to Tract Resources

- Understanding Census Geography Hierarchies: https://www.census.gov/programs-surveys/geography/guidance/hierarchy.html
- Understanding Census Geography IDs: https://www.census.gov/programs-surveys/geography/guidance/geo-identifiers.html

Appendix D. Geocoding Best Practices

Imputation

1. When address level geocodes cannot be obtained, imputation may be used for records geocoded to areas that cross county boundaries.
2. Imputation to county level is encouraged but not required.
3. For areas with similar place names a second geographic identifier should be used to distinguish them (e.g., ZIP code).
4. Imputation should be weighted by population and, when possible, using the same demographic characteristics of the data being imputed (e.g., female breast cancer should not include males or children in the population weighting).
5. CDC will accept decimal place to a 10^{th} for a count and will display whole numbers. Fractions will be kept to calculate county totals of proper amount.

Appendix E. Acknowledgments

The Tracking Program thanks Wes Kortuem (Arizona Tracking Program) and Angie Werner and Craig Kassinger (CDC Tracking Program) for compiling the information in this document. The Tracking Program also thanks the following members of the Tracking Program Standards and Network Development Geospatial Team who contributed to the creation of the Geocoding Standards and have provided ongoing input:

Arizona Department of Health Services: Josue Barboza, Nicole Eiden, David Olsen, Wes Kortuem, Marla Kostuk; **Centers for Disease Control and Prevention**: Craig Kassinger, Aaron Grober, Bob Kennedy, Michele Monti, TJ Pierce, Meekie Shin, Jen Shriber, Aaron Vinson, Patrick Wall, Angela Werner, Sam Wotiz; **Colorado Department of Public Health & Environment**: Kevin Berg, Ben White, Devon Williford; **Connecticut Department of Public Health**: Gary Archambault, Patricia Przysiecki; **Delaware Health and Social Services**: Tabatha Offutt-Powell; **Florida Department of Health**: Chris DuClos, Jessi Joiner, Keshia Reid; **Green River**: Michael Knapp; **Kansas Department of Health and Environment**: Henri Menager; **Louisiana Department of Health**: John Anderson, Kathleen Aubin, Kate Friedman, Adrian Savella, Alexis Williams; **Massachusetts Department of Public Health**: Glennon Beresin, Erin Collins, Braden Miller; **Maine Department of Health and Human Services**: Jessica Bonthius, Kathy Decker, Rebecca Lincoln, Lisa Parker, Chris Paulu; **Michigan Department of Health & Human Services**: Jill Maras, Sydney Ogden; **Minnesota Department of Health**: Blair Sevcik, Jessie Shmool; **Missouri Department of Health and Senior Services**: Kathleen Kloeppel, Jeff Patridge, Scott Patterson, Jen Weaver; **NAACCR**: Recinda Sherman; **NAHDO**: Charles Hawley; **NAPHSIS**: Kristin Simpson, Shae Sutton; **New Hampshire Department of Health & Human Services**: Katie Bush, Samuel Harris, Dennis Holt, Jennifer Howley, Jessica Sagona, Nicholas Shonka; **New Jersey Department of Health**: Barb Goun; **New Mexico Department of Health**: Tony Fristachi, Lois Haggard; **New York City Department of Health and Mental Hygiene**: Grant Pezeshki; **New York State Department of Health**: Jeff Bryant, Douglas Done, Tabassum Insaf, Sanjaya Kumar, Neil Muscatiello, Arjita Raj, Abby Stamm; **Oregon Health Authority**: Eric Main; **Rhode Island Department of Health**: Peter DiPippo, Joseph Maya-Rodriguez, Jay Metzger, Catherine Schultz, Mike Simoli; **Ross Strategic**: Mary Byrne, Jessie Doody, Samer Khan, Jen Major, Lissette Palestro; **Texas A&M**: Payton Baldridge, Dan Goldberg; **Utah Department of Health**: Johnny Auld, Sam LeFevre, Nelson Long, Matt McCord; **Vermont Department of Health**: Daniel Jarvis, Pete Young; **Washington State Department of Health**: Chris Ahmed, Lauren Freelander, Lillian Morris; **Wisconsin Department of Health Services**: Jenny Camponeschi, Paul Creswell.